

Development of bisulfite-free DNA methylation sequencing methods and their application to cell-free DNA for cancer detection



Paulina Siejka-Zielińska

Kellogg College

Ludwig Institute for Cancer Research

Nuffield Department of Medicine

University of Oxford

This thesis is submitted in fulfilment of the requirements for the degree of
Doctor of Philosophy in Clinical Medicine

Trinity 2020

This thesis is submitted to the Department of Medical Science, University of Oxford, in fulfilment of the requirements for the degree of Doctor of Philosophy. This thesis is the result of my own work unless otherwise stated and does not constitute part of any other thesis.

Word count: 49032

Paulina Siejka-Zielińska, Kellogg College

Development of bisulfite-free DNA methylation sequencing methods and their application to cell-free DNA for cancer detection.

Paulina Siejka-Zielińska

Doctor of Philosophy in Clinical Medicine

Kellogg College

Trinity 2020

Abstract

5-Methylcytosine (5mC) is the most abundant epigenetic mark in DNA regulating gene expression and influencing crucial cellular processes. Aberrant DNA methylation is involved in pathogenesis of human malignancies. Detection of cancer-specific methylation patterns in cell-free DNA (cfDNA) is a promising approach for development of minimally invasive, liquid biopsy assays. Most of methylation sequencing methods are based on bisulfite treatment that causes an extensive DNA damage, therefore are not ideal for analysis of minute amount of cfDNA. Unfortunately, technological limitations affect whole-genome methylation analysis of cfDNA. During the course of this project, we developed and validated novel approaches addressing the most important drawbacks of currently available technologies.

First, we developed and validated enrichment-based cfMeSeal approach by combining enzymatic and chemical labelling for cost-effective 5mC sequencing in cfDNA.

Second, we designed and developed a bisulfite-free method, Tet-assisted Pyridine Borane Sequencing (TAPS) for base-resolution 5mC sequencing. We demonstrated that TAPS is a mild reaction and outperforms bisulfite sequencing (BS-seq) at many aspects including improved sequence quality, better mapping statistics and direct conversion of modified sites.

Finally, we confirmed that TAPS can be applied for cfDNA methylation sequencing and provides comprehensive information of cfDNA methylation and fragmentation. Information

gained from cfDNA TAPS sequencing enables detection of hepatocellular carcinoma and pancreatic cancer with high accuracy.

Together, we demonstrated that newly developed sequencing tools can outperform existing technologies and facilitate comprehensive cfDNA methylation sequencing and provide multidimensional information about cfDNA with high cancer predictive potential.

This thesis is dedicated to

My husband, Marcin,

My brother, Sebastian,

And my parents.

*In ordinary life, we hardly realize
that we receive a great deal more than we give,
and that it is only with gratitude that life becomes rich.*

Dietrich Bonhoeffer

Acknowledgments

I would like to thank all people that helped me during 4 years of my DPhil project.

First, I would like to express my gratitude to my supervisors: my primary supervisor Dr Chunxiao Song, my secondary supervisor Professor Skirmantas Kriaucionis and my former secondary supervisor Professor Panagis Filippakopoulos for invaluable scientific advices, inspiring discussions and guiding me thorough this project. I thank Ludwig Cancer Research and Kellogg collage for giving me the opportunity to study in Oxford and for funding my studies.

I would also like to thank all people involved in this project, especially Dr Yibin Liu for fruitful and collaborative work on TAPS development, Dr Jingfei Cheng, Gergana Velikova, Felix Jackson and Dr Benjamin Schuster-Böckler for guidelines on data analysis part and helping me to develop computational skills and Dr Anandhakumar Chandran for rrTAPS development. I wish to acknowledge all people involved in collecting clinical samples and recruiting patients, especially Prof Ellie Barnes, Dr Shivan Sivakumar, Dr Emma Culver, Luminita Puta, Zahir Soonawalla, Srikanth Reddy, Mike Silva and Dr Noor Bekkali for their help and medical expertise as well as all patients and healthy volunteers that donated their blood for this study for their courage and sacrifice.

I would like to acknowledge all current and former members of Song and Kriaucionis labs. My special thanks go to Michael, Felix and Gergana for helping me to proofread this dissertation as well as to other lab members: Ying, Masato, Xiufei, Jinfeng, Lei, Fang, Lin, Sophie, Paolo,

Olena, Ashvina, Hiromi, Mankgopo, Marketa, Wenjun and all Oxford friends for collaborative work on other research projects, all formal and informal discussions, support and friendship. Finally, my greatest thanks go to my family for their never-ending love and support. The most special thanks go to my beloved husband, Marcin, for sharing all good and bad moments of my DPhil, amazing morning coffee and helping me to find balance and enjoy time outside the lab. Owing to his support, motivation and faith in me I found strength and courage to start and complete this DPhil. I would like to thank my brother, Sebastian, for his unique sense of humour that always makes me smile and for sharing favourite songs, so I have not lost my teenage spirit yet. Last but not least, I thank my parents for helping me to grow up and become the person I am now.

Contents

1	Introduction.....	1
1.1	Epigenetics.....	1
1.1.1	Introduction to epigenetics.....	1
1.1.2	Cytosine modifications and their role in epigenetic regulation	3
1.1.3	Biological role of DNA methylation in healthy and cancer	12
1.2	DNA methylation for non-invasive cancer detection.....	19
1.2.1	Significance of early cancer diagnosis	19
1.2.2	Liquid biopsy for cancer diagnosis and detection.....	22
1.2.3	Cell-free DNA as cancer biomarker	25
1.2.4	cfDNA methylation signatures in liquid biopsy	30
1.2.5	Diagnosis and management of pancreatic cancer	34
1.2.6	Diagnosis and management of hepatocellular carcinoma	38
1.3	Overview of epigenetic tools for DNA methylation analysis.....	41
1.3.1	Global analysis of DNA methylation	42
1.3.2	Enrichment-based methods.....	44
1.3.3	Base-resolution methods for whole-genome methylation analysis.....	50
1.3.4	Methylation arrays.....	54
1.3.5	Third generation sequencing methods for methylome analysis.....	55
1.3.6	Concluding remarks	57
2	Research objectives	59

3	Cell-free DNA MeSeal method for 5mC enrichment.....	61
3.1	Introduction.....	61
3.2	Objectives	62
3.3	General design of cfMeSeal method	63
3.4	Non-mammalian, recombinant TET enzymes can be successfully expressed and purified from <i>E. coli</i>	65
3.5	5mC in cfDNA can be enriched by chemical labelling	68
3.6	cfMeSeal sequencing provides specific methylation information	72
3.7	cfMeSeal sequencing captures cancer-specific methylation patterns in cfDNA	76
3.8	Mutated NgTET1 does not improve 5mC labelling reaction stability	82
3.9	Discussion	86
3.9.1	cfDNA methylation patterns in cancer	86
3.9.2	Limitations	89
3.9.3	Concluding remarks	91
3.10	Materials and Methods	92
3.10.1	Preparation of spike-in control oligos	92
3.10.2	Expression and purification of NgTET1	92
3.10.3	Expression and purification of ccTET.....	93
3.10.4	Preparation of NgTET1 A212V	94
3.10.5	TET activity test.....	94
3.10.6	Optimisation of UDP-glucose concentration for 5hmC blocking.....	95
3.10.7	Optimisation of one-pot blocking/labelling reaction on model DNA.....	95

3.10.8	Comparison of 5mC labelling reaction with NgTET1 and ccTET	95
3.10.9	Plasma Samples Collection and Processing.....	96
3.10.10	5-methylcystosine pull-down in cfDNA.....	97
3.10.11	5mC pull-down validation by PCR	98
3.10.12	Data processing	99
3.10.13	Differential methylation analysis	99
4	Development of Tet-assisted Pyridine Borane Sequencing (TAPS) for base-resolution DNA methylation analysis.....	100
4.1	Introduction.....	100
4.2	Motivation.....	102
4.3	Borane compounds reduce 5caC to DHU.....	102
4.4	Recombinant TET enzymes for TAPS	106
4.5	Spike-in controls for 5mC conversion monitoring.....	108
4.6	5mC conversion can be monitored by restriction digestion assay	109
4.7	Pic-borane reduces oxidised 5caC in a genome.....	110
4.8	mTet1 oxidation followed by pyridine borane reduction improves 5mC detection in a genome	117
4.9	TAPS exhibits improved sequencing quality compared to BS-seq and provides reliable whole-genome, base-resolution methylation landscape in mESC.....	121
4.10	Discussion	129
4.10.1	Production of recombinant TET enzymes	129
4.10.2	Methylation analysis by TAPS	131

4.10.3	Concluding remarks	133
4.11	Materials and Methods	133
4.11.1	Preparation of DNA oligonucleotides for MALDI	133
4.11.2	Preparation of spike-in DNA controls for TAPS	134
4.11.3	Preparation of 222 bp model DNA	135
4.11.4	DNA digestion and HPLC–MS/MS analysis.....	136
4.11.5	Expression and purification of mTet1CD.....	137
4.11.6	TET activity assay	137
4.11.7	NgTET1 oxidation.....	138
4.11.8	mTet1CD oxidation.....	138
4.11.9	Validation of TAPS conversion with Taq α I assay.....	139
4.11.10	TAPS on model DNA.....	139
4.11.11	Preparation of Illumina multiplexing adapters	139
4.11.12	mESC culture and genomic DNA preparation.....	140
4.11.13	Pic-borane TAPS (in MeOH) sequencing of mESC genomic DNA.....	140
4.11.14	TAPS optimisation with pic-BH ₃	141
4.11.15	Pic-borane TAPS sequencing of mESC genomic DNA.....	142
4.11.16	Low-input whole-genome TAPS (pic-BH ₃) with ssDNA library kit.....	142
4.11.17	Cell-free DNA TAPS (pic-BH ₃)	143
4.11.18	TAPS with pyridine borane – initial conditions.....	143
4.11.19	TAPS optimisation with pyridine borane	144

4.11.20	TAPS condition with pyridine borane – second screening sequencing	145
4.11.21	WGBS.....	146
4.11.22	DNA damage test on fragments with different length	147
4.11.23	WGBS data processing.....	147
4.11.24	TAPS data pre-processing	148
4.11.25	Sequencing quality analysis of TAPS and WGBS.....	148
4.11.26	Coverage analysis of TAPS and WGBS.....	148
4.11.27	Analysis of cytosine modifications measured by TAPS and WGBS.....	149
5	Methylation profiling in cell-free DNA by TAPS – proof-of-concept study	150
5.1	Introduction.....	150
5.2	Motivation.....	151
5.3	TAPS provides reliable, base-resolution cfDNA methylation landscape.....	152
5.4	CfDNA fragmentation profiles are preserved in TAPS and have predictive potential	176
5.5	Discussion.....	181
5.6	Concluding remarks.....	187
5.7	Materials and Methods	188
5.7.1	Preparation of carrier DNA and spike-in controls	188
5.7.2	Preparation of mTET1CD.....	188
5.7.3	Preparation of sequencing adapters.....	188
5.7.4	mTet1CD oxidation.....	188
5.7.5	Pyridine borane reduction.....	189

5.7.6	Preparation of low input TAPS libraries	189
5.7.7	Collection and preparation of cfDNA samples	190
5.7.8	cfDNA TAPS with carrier DNA	191
5.7.9	TAPS mapping and pre-processing.....	191
5.7.10	Analysis of cfDNA methylation by TAPS	192
5.7.11	Differential methylation analysis	193
5.7.12	Fragmentation analysis	194
5.7.13	Model building	194
6	Endonuclease-enrichment TAPS.....	196
6.1	Introduction.....	196
6.2	Objectives	197
6.3	DHU is cleaved by restriction nucleases	198
6.4	TAPS method can be adopted to reduce-representation and endonuclease enrichment protocols.....	200
6.5	eeTAPS provides reliable methylation information.....	202
6.6	eeTAPS provides comprehensive information about methylation distribution.....	206
6.7	eeTAPS can be successfully applied for low input applications	209
6.8	eeTAPS is cost-effective approach for methylation sequencing.....	210
6.9	Discussion	211
6.10	Concluding remarks.....	215
6.11	Materials and Methods	215
6.11.1	Preparation of spike-in controls	215

6.11.2	Preparation of carrier DNA.....	216
6.11.3	Expression and purification of mTet1CD.....	216
6.11.4	mESCs culture and isolation of genomic DNA	216
6.11.5	mTet1CD oxidation.....	216
6.11.6	eeTAPS	216
6.11.7	rrTAPS.....	217
6.11.8	Screening for DHU digesting endonucleases	218
6.11.9	Data analysis for eeTAPS.....	219
6.11.10	Data analysis for rrTAPS.....	219
6.11.11	Comparison of wgTAPS, eeTAPS and rrTAPS in mESC.....	219
7	Discussion	221
8	Conclusions.....	232
9	References.....	234
10	Supplementary Materials	312
10.1	Appendix 1 - Glossary.....	312
10.2	Appendix 2 - Publications.....	318
10.3	Appendix 3 – Supplementary Materials to Chapter 3	319
10.4	Appendix 4 – Supplementary Materials to Chapter 4	327
10.5	Appendix 5 – Supplementary Materials to Chapter 5	331
10.6	Appendix 6 – Supplementary Materials to Chapter 6	338

*As you set out for Ithaka
hope the voyage is a long one,
full of adventure, full of discovery.*
C. P. Cavafy Ithaca

1 Introduction

1.1 Epigenetics

1.1.1 Introduction to epigenetics

Although the discovery of the structure of DNA by Watson and Crick (Watson and Crick, 1953) and sequencing of the human genome almost 50 years later (Lander *et al.*, 2001) were great advantages for genetics and biology, DNA sequence itself failed to explain cellular heterogeneity observed in metazoans and how the cells within one organism can develop completely different phenotypes despite sharing the same genome. This phenotypic diversity is the main interest of epigenetics which is focused on the discovery and understanding of processes that translate genotype into phenotype. From a modern perspective, epigenetics can be defined as a set of DNA sequence-independent mechanisms that govern transcriptional programs in a cell and in this way regulate read-out of the genome and shape cell phenotype (Rivera and Ren, 2013).

Epigenetic mechanisms include chemical modifications on DNA and histone proteins as well as other processes such as gene silencing by non-coding RNA (ncRNA). In mammalian cells, genetic material is stored as nucleoprotein complex. DNA string of 147 bp is wrapped around histone octamer consisting of stoichiometric assembly of histones H2A, H2B, H3 and H4 to form the basic unit of chromatin - nucleosome. Nucleosomes can be further folded to create more condensed forms: chromatin fibres, chromosome domains and chromosomes (Luger *et al.*, 1997; Rivera and Ren, 2013; Bonev and Cavalli,

2016). Additional chemical modifications can be deposited on DNA and histone proteins by enzymes and these modifications play an important role in transcriptional regulation (Leung *et al.*, 2015; Liu, Denizio and Kohli, 2016). Epigenetic marks on histone proteins and DNA act to reinforce complex regulatory networks that modulate chromatin structure and accessibility and recruit a broad range of chromatin-interacting proteins. Importantly, epigenetic marks are plastic and reversible, as they can be written or erased from specific parts of the genome, providing unique and dynamic ways to repress or activate specific genes when required. In this way epigenetics has a profound effect on gene expression, governing transcriptional programs and, more broadly, regulating key biological processes (Borgel *et al.*, 2010; Rivera and Ren, 2013; Leung *et al.*, 2015; Klemm, Shipony and Greenleaf, 2019).

Histone modifications are an important part of epigenetic regulation. Histone proteins consist of a globular core domain and positively charged N-terminal tails. Histone tails serve as a template for post-translational modifications and epigenetic marks on histone tails play an important role in regulation of interactions between nucleosomes and chromatin structure (Jenuwein and Allis, 2001; Bannister and Kouzarides, 2011).

Numerous types of histone marks and their involvement in regulation of the genome have been described in the literature, with histone lysine acetylation and methylation being among the best studied (Bernstein *et al.*, 2005, 2006; Creyghton *et al.*, 2010; Tan *et al.*, 2011; Tian *et al.*, 2012; Farrelly *et al.*, 2019). Histone acetylation is predominantly linked to accessible, transcriptionally active chromatin (Bernstein, Meissner and Lander, 2007; Leung *et al.*, 2015; Raisner *et al.*, 2018). Acetylation can partially neutralise positive charge of histone tails and in this way attenuate their interactions with DNA, which results in an increase in chromatin accessibility and facilitates transcription. Additionally, histone acetylation marks such as the well-established marker of active enhancers,

H3K27ac3, directly participate in recruiting and binding of transcriptional machinery (Creyghton *et al.*, 2010; Raisner *et al.*, 2018).

In contrast, histone lysine methylation has a more diverse effect on gene expression regulation (Bernstein *et al.*, 2005, 2006; Akkers *et al.*, 2009; Maunakea *et al.*, 2010; Okitsu, Hsieh and Hsieh, 2010; Leung *et al.*, 2015; Gates, Foulds and O'Malley, 2017). For instance, H3K4me3 enriched in active promoters is associated with active transcription and RNA polymerase II occupancy as well as it plays a role in recruiting regulatory proteins and transcriptional factors to the transcription site and formation of pre-initiation complex (Bernstein *et al.*, 2005; Okitsu, Hsieh and Hsieh, 2010; Gates, Foulds and O'Malley, 2017). On the contrary, H3K27me3 mark is mostly associated with transcriptionally inactive, Polycomb-repressed regions (Leung *et al.*, 2015).

Apart from histone marks, DNA modifications, especially cytosine methylation and hydroxymethylation, significantly contribute to epigenetic regulation of the genome (Smith and Meissner, 2013). Modified cytosines interact with histone marks to establish and maintain the transcriptional status of specific genes (Schultz *et al.*, 2015; Argelaguet *et al.*, 2019). Since DNA methylation is the main focus of my DPhil project, in the following subsections I will introduce the origin and biological functions of cytosine modifications and explain their significance in physiological processes and cancer.

1.1.2 Cytosine modifications and their role in epigenetic regulation

5-Methylcytosine (5mC) is the most common epigenetic mark in mammalian DNA. It was first discovered in early 1900s (Johnson and Coghill, 1925), however, it was not until 1948 when its existence was finally confirmed and widely accepted by the scientific community (Hotchkiss, 1948). Since then 5mC has been the most extensively studied epigenetic modification in DNA. The presence of 5mC was confirmed among various

species including bacteria, plants and mammals, however, its biological function and genomic distribution are highly distinct between these groups (Law and Jacobsen, 2010; Mohapatra and Biondi, 2017). In this dissertation I will focus exclusively on DNA methylation in mammals.

In mammalian cells DNA methylation exists mainly in symmetric CG dinucleotides known as CpG sites, however, some cells and tissues including brain and stem cells exhibit a relatively high level of 5mC in non-CG context (Lister *et al.*, 2009, 2013; Law and Jacobsen, 2010; Xie *et al.*, 2012; Zhang and Pradhan, 2014; Schultz *et al.*, 2015). DNA methylation is enzymatically introduced to the genome by the transfer of methyl group from S-Adenosyl-L-methionine (SAM) to the fifth position of cytosine in a reaction catalysed by DNA methyltransferases (DNMTs) (Figure 1-1). Methylation marks can be deposited directly on unmethylated DNA by *de novo* methyltransferases DNMT3A and DNMT3B (Okano, Xie and Li, 1998) or they can be maintained by maintenance methyltransferase, DNMT1, which recognizes hemimethylated CpG sites and restores them to a fully methylated state (Hermann, Goyal and Jeltsch, 2004).

5mC exists in relatively high abundance in mammalian genomes and it is estimated that around 5% of all cytosines in the genome are methylated (Breiling and Lyko, 2015). Approximately 70-80% of all CpG sites in the genome are methylated and methylation can be found in a variety of regions across the whole genome. However, methylated sites are depleted in specific genomic features such as CpG islands (CGIs), enhancers and active promoters (Lister *et al.*, 2009; Ficz *et al.*, 2011; Stadler *et al.*, 2011).

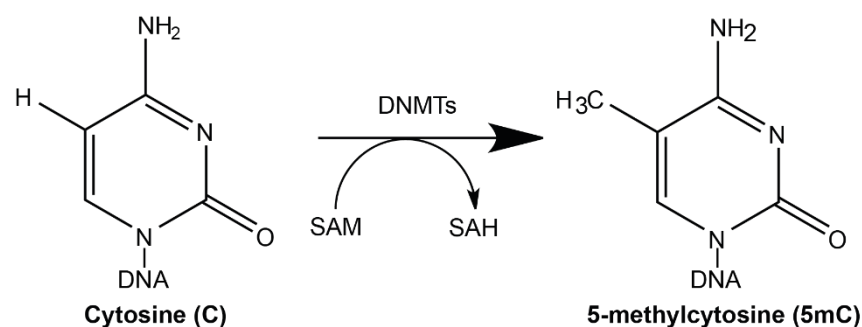


Figure 1-1. Reaction of cytosine methylation catalysed by DNMT proteins. DNMTs catalyse transfer of methyl group from SAM cofactor to fifth position of cytosine in DNA to produce 5mC. As a results of methyl group donation SAM is converted to S-Adenosyl-L-homocysteine (SAH).

DNA methylation plays an important role in gene expression regulation via two mechanisms. First, 5mC can directly regulate binding of transcriptional factors (TFs) (Yin *et al.*, 2017). Second, it can recruit chromatin binding proteins and initiate changes in chromatin structure with further impact on gene expression (Feng and Zhang, 2001). In general, 5mC in promoter regions is considered as a repressive mark, inhibiting TFs binding, blocking transcription initiation and promoting closed chromatin structure (Feng and Zhang, 2001; Borgel *et al.*, 2010; Ficiz *et al.*, 2011; Han *et al.*, 2011; Yang *et al.*, 2020). It has been shown that low DNA methylation and presence of activating histone marks are more prevalent in the regulatory regions of highly expressed genes while high DNA methylation and repressive histone marks are associated with lowly expressed or silenced genes (Maunakea *et al.*, 2010; Leung *et al.*, 2015). On the other hand, some studies suggest that the role of DNA methylation may not be limited only to transcriptional repression (Lister *et al.*, 2009; Han *et al.*, 2011). Although promoter methylation is typically associated with transcriptional repression (Han *et al.*, 2011), hypermethylation of gene bodies has been observed in highly transcribed genes, suggesting that the effect of 5mC on gene expression may be location-dependent or modified by other epigenetic mechanisms (Lister *et al.*, 2009).

Similarly, DNA methylation may have a distinct effect on TFs binding (Stadler *et al.*, 2011; S. Hu *et al.*, 2013; Yin *et al.*, 2017; Yang *et al.*, 2020). While binding of some TFs is inhibited by methylation, some studies have shown that a fraction of TFs can preferentially bind to methylated cytosines, promoting chromatin changes and activating expression of target genes. Interestingly, despite the well-established role of 5mC in establishing closed chromatin state via recruitment of MeCP1 protein complex and histone deacetylation (Feng and Zhang, 2001), other studies have suggested that TFs binding to 5mC can induce local methylation loss and promote a more active chromatin state (Stadler *et al.*, 2011; Yin *et al.*, 2017).

Although the main role of 5mC is transcriptional repression, the above-mentioned examples suggest that apart from its canonical repressive role DNA methylation may also contribute to gene expression activation. The final impact of 5mC on genome regulation is a cumulative effect of multiple factors including location in the genome, surrounding histone marks, TF occupancy and biological state of the cell. This context-dependent effect reflects the dynamic character of epigenetic regulation.

Another important feature of epigenetic regulation is reversibility. It means that epigenetic marks can be removed from the genome or replaced by others. Despite the fact that the mechanism of 5mC deposition in the genome had been known for years, little was known about its removal. For a long time, loss of 5mC was attributed to passive demethylation that occurs during cell divisions (Kohli and Zhang, 2013; Wu and Zhang, 2014).

Passive demethylation refers to 5mC loss that could be observed during DNA replication. Each replication round results in generation of hemimethylated CpG sites since the newly synthesised strand is unmodified. Hemimethylated CpG sites are usually methylated by DNMT1 enzyme and the fully methylated state is maintained. In the absence of effective

DNA methylation maintenance, symmetric CG methylation is not restored and after subsequent replication rounds, the DNA methylation level decreases as a result of passive dilution (Figure 1-2).

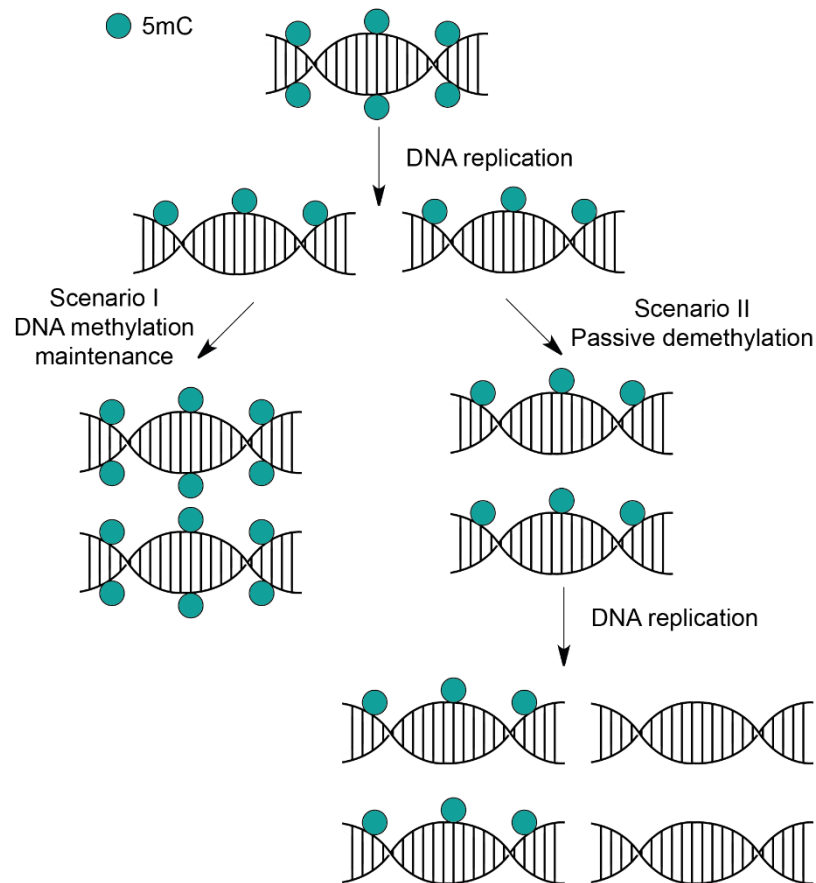


Figure 1-2. Overview of passive demethylation pathway. In the absence of DNA methylation maintenance, fully methylated state of CpG sites in hemimethylated DNA molecules generated after replication cannot be restored and methylation is passively lost during subsequent replication rounds.

On the other hand, passive demethylation does not explain the global demethylation wave occurring independently of replication in the paternal pronucleus and in Primordial germ cells during early development (Hajkova *et al.*, 2010). At the beginning of 2000s, computational analysis of Fe(II)/ α -ketoglutarate (α -KG)-dependent oxygenases, JBP1 and JBP2, involved in thymine oxidation and formation of glucosylated 5-hydroxyuracil (Base J) in *Trypanosoma brucei* led to identification of Tet-eleven translocation (TET)

proteins as their mammalian paralogues (Iyer *et al.*, 2009) and initiated further investigation of these enzymes.

Subsequent experimental studies demonstrated that TET enzymes are not only active dioxygenases catalysing oxidation of methylated cytosines in the genome but also confirmed the existence of an active DNA demethylation pathway (Tahiliani *et al.*, 2009; He *et al.*, 2011; Ito *et al.*, 2011).

TET proteins catalyse stepwise oxidation of 5mC to 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC) in DNA in a Fe(II)/ α -KG-dependent manner (Tahiliani *et al.*, 2009; He *et al.*, 2011; Ito *et al.*, 2011). Final oxidation products, 5fC and 5caC, can be subsequently removed from the genome by thymine DNA glycosylase (TDG) and replaced with unmodified cytosine via base excision repair (BER) pathway (Figure 1-3).

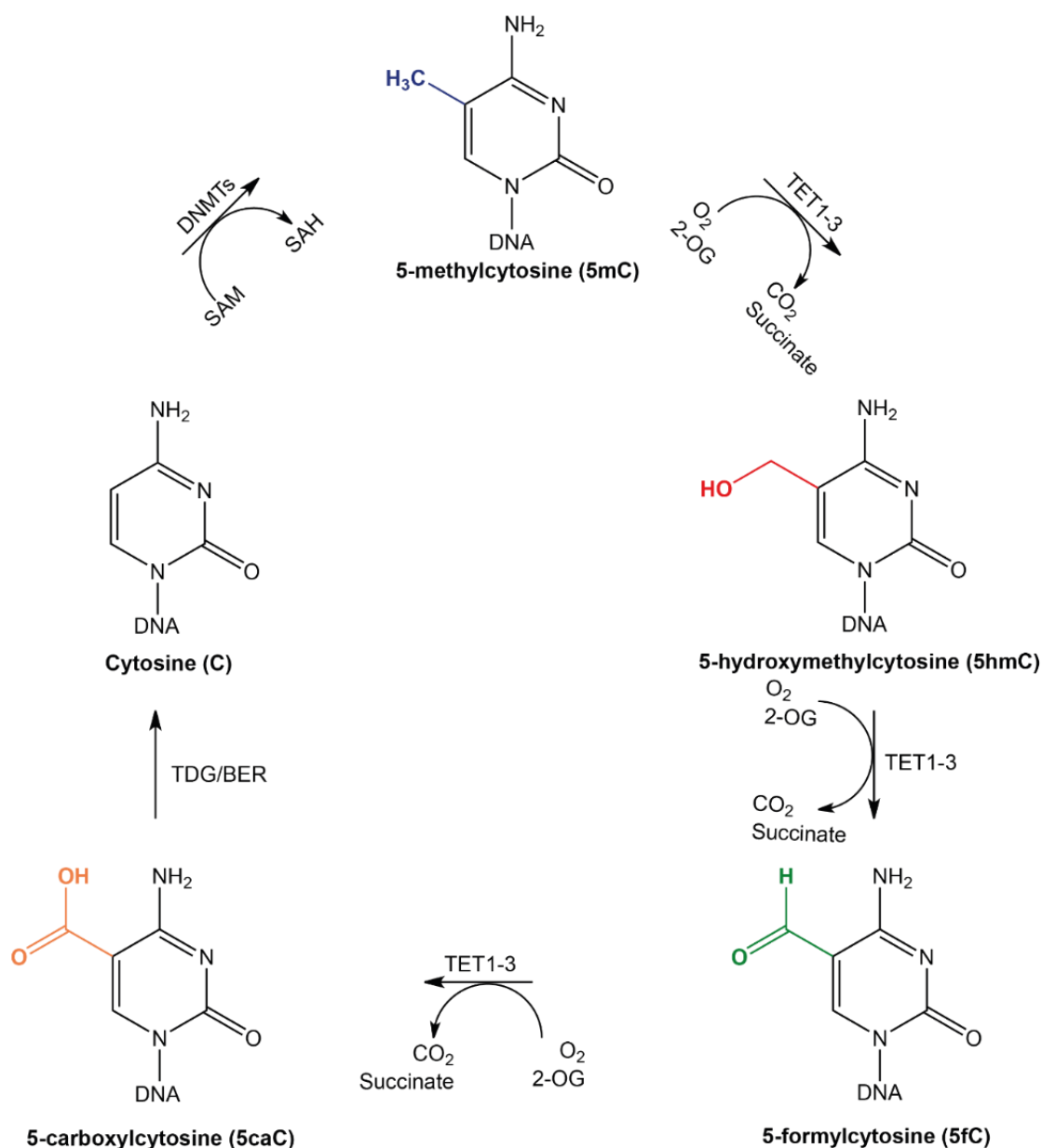


Figure 1-3. Active DNA demethylation pathway and generation of oxidised cytosine variants. 5mC in DNA can be oxidised by TET enzymes to 5caC, which can be then removed from the genome and replaced by unmodified cytosine via TDG/BER. Apart from being intermediates in demethylation pathway, oxidised cytosines (5hmC, 5fC and 5caC) may also act as stable epigenetic marks and play regulatory functions in the genome.

Owing to their involvement in active demethylation pathway, TET proteins are considered as important epigenetic regulators. The mammalian TET family comprises of three enzymes: TET1, TET2 and TET3 (Figure 1-4) (Tahiliani *et al.*, 2009; He *et al.*, 2011; Ito *et al.*, 2011). TET-mediated DNA demethylation is involved in regulation of key biological processes including early development, transcription activation,

establishment of bivalent promoters, regulation of enhancers and cellular self-renewal (Ito *et al.*, 2010; Iqbal *et al.*, 2011; Shen *et al.*, 2013; Kong *et al.*, 2016). Moreover, products of active DNA oxidation themselves serve as stable epigenetic marks playing regulatory functions in the genome (Ficz *et al.*, 2011; Song *et al.*, 2011; Kellinger *et al.*, 2012; Iurlaro *et al.*, 2013; Shen *et al.*, 2013).

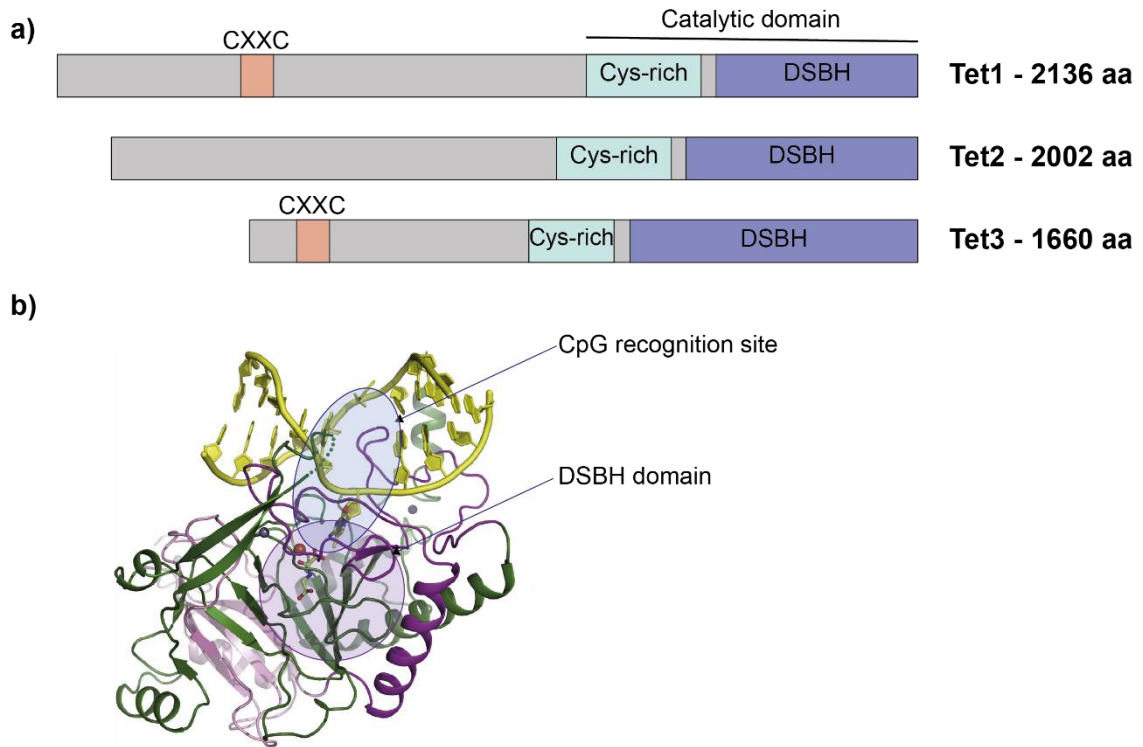


Figure 1-4. Human TET proteins. (a) Structure of TET proteins. CXXC domain is involved in DNA binding and has high affinity to CpG sites. Tet2 lacks CXXC domain. Catalytic domains of TETs consists of Cystein-rich region and double-stranded β -helix domain (DSBH). DSBH domain forms catalytic core and interacts with Fe(II) and 2-OG. Adopted from An, Rao and Ko, 2017. (b) Crystal structure of hTET2 catalytic domain. Cystein-rich domain is coloured in purple and DSBH domain is coloured in green. CpG recognition site and DSBH catalytic core are marked with circle. Adopted from L. Hu *et al.*, 2013.

5hmC is the product of first step of 5mC oxidation and is the second most abundant cytosine modification in the genome. It is estimated that around 0.1-1% of all cytosines are hydroxymethylated depending on cell type with particularly high levels in stem cells and Purkinje neurones (Kriaucionis and Heintz, 2009; Yu *et al.*, 2012; Lister *et al.*, 2013). Distribution of 5hmC in the genome differs significantly from DNA methylation. While CpG methylation exhibits bimodal distribution with most sites highly modified or

completely unmodified, hydroxymethylated CpG sites show low modification level around 20-30%. Moreover, contrary to DNA methylation which is usually symmetric, most 5hmC in CG context exhibits strand asymmetry (Yu *et al.*, 2012).

DNA hydroxymethylation is enriched in specific genomic regions including gene bodies and distal regulatory elements. 5hmC is mainly associated with euchromatin and its presence in promoter regions is frequently accompanied by activating histone marks and higher expression of target genes. Notably, promoters enriched in both 5mC and 5hmC exhibit higher expression of target genes compared to methylated promoters, which suggest that 5hmC can at least partially neutralise the repressive effect of 5mC (Ficz *et al.*, 2011; Song *et al.*, 2011).

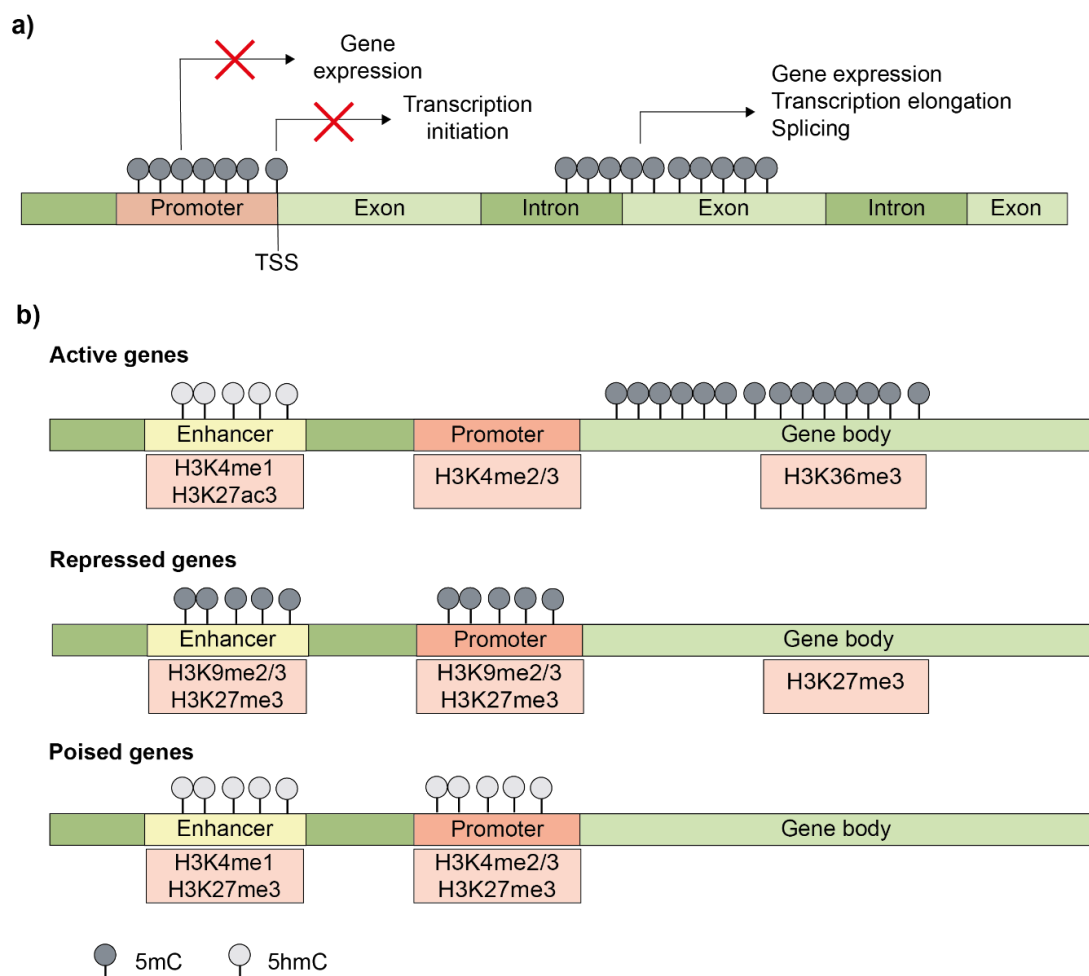


Figure 1-5. Simplified overview of effect of DNA methylation and hydroxymethylation on gene expression. (a) Location-dependent effect of DNA methylation on gene expression. Methylation of promoter regions and TSS is associated with transcriptional repression while methylation in

gene body plays a role in gene expression, transcription elongation and splicing. (b) Distinct epigenetic landscape of active, repressed and poised genes. Distribution of 5mC, 5hmC and key histone marks in promoters, enhancers and gene bodies differ depending on transcriptional status of the gene.

Another two products of 5mC oxidation, 5fC and 5caC, exist at very low levels in the genome, estimated to be respectively 10-fold and 100-fold lower than 5hmC (Ito *et al.*, 2011). Both 5fC and 5caC are direct substrates for TDG enzyme and for a long time they were considered only as intermediates of DNA demethylation. However, according to recent studies they contribute to epigenetic regulation. While 5fC and 5caC are actively removed from bivalent promoters and distal regulatory elements via TET/TDG-dependent manner, they exhibit surprisingly conserved distribution in proximal regulatory elements (Shen *et al.*, 2013). Moreover, some studies suggest that 5fC and 5caC can be recognized by specific regulatory proteins or affect function of RNA polymerase II with direct effect on transcription (Kellinger *et al.*, 2012; Iurlaro *et al.*, 2013; Spruijt *et al.*, 2013; Wang *et al.*, 2015).

The fact that DNA demethylation pathway serves both to remove epigenetic marks and to generate novel DNA modifications with unique and specific regulatory functions adds an additional layer of complexity to epigenetic regulation at the DNA level. However, existence of such complexity in mammals is justified by biological necessity. All mammalian cells depend on activation or deactivation of transcriptional programs that allow them to perform cell-specific functions, pass another developmental stage or respond to the external stimuli. Together with other epigenetic factors, DNA methylation and demethylation are an important part of this regulatory network.

1.1.3 Biological role of DNA methylation in healthy and cancer

Correct deposition, maintenance and erasure of DNA methylation are crucial for establishment and maintenance of physiological cellular functions. DNA methylation and

hydroxymethylation are known to be involved in regulation of early development (Lister *et al.*, 2013; Argelaguet *et al.*, 2019), pluripotency (Ito *et al.*, 2010; Ficz *et al.*, 2011), genomic imprinting (Kaneda *et al.*, 2004) and genome stability (Kim *et al.*, 2004).

During early development, DNA methylation and demethylation are important factors contributing to chromatin remodelling, establishment of developmental pathway and cell-lineage commitment. Upon fertilization, paternal DNA undergoes global TET3-dependent demethylation, while maternal DNA is protected against active demethylation by Stella protein and loses methylation in a passive way during cell divisions. Another wave of global, active demethylation occurs during Primordial germ cell development (Mayer *et al.*, 2000; Santos *et al.*, 2002; Hajkova *et al.*, 2010; Iqbal *et al.*, 2011). During implantation and lineage specification, *de novo* methylation is introduced into promoter and enhancer regions to repress genes related to pluripotency and lineage-specificity. In parallel, demethylation of specific regulatory regions occurs in cells undergoing final differentiation. In this way gain and loss of DNA methylation is involved in establishment of cell-specific transcriptional programs and cellular identity (Borgel *et al.*, 2010; Argelaguet *et al.*, 2019). Disruption of both DNA methylation and demethylation has a dramatic effect on development and leads to embryonic lethality and developmental abnormalities (Li, Bestor and Jaenisch, 1992; Okano *et al.*, 1999; Kaneda *et al.*, 2004; Gu *et al.*, 2011).

DNA methylation is also involved in establishment of genomic imprinting and X chromosome inactivation (XCI) (Tucci *et al.*, 2019). Imprinted genes are expressed only from one parental allele while the second allele is transcriptionally inactive. Imprinted status is defined early in development (Tucci *et al.*, 2019). DNMTs and DNA methylation are responsible for establishment of genomic imprinting (Li, Beard and Jaenisch, 1993; Bourc'h's *et al.*, 2001; Hata *et al.*, 2002; Kaneda *et al.*, 2004; Matsuzaki *et al.*, 2018).

DNA methylation marks imprinting control regions in the genome and directs further methylation of target genes. Depletion of DNMTs followed by loss of methylation in imprinted genes allows for reactivation of their expression and leads to developmental defects.

DNA methylation is also important for XCI process. In female mammals one copy of the X chromosome is silenced by epigenetic mechanisms involving DNA methylation, long ncRNAs and histone marks. Abnormalities of X chromosome inactivation can be associated with developmental defects and human diseases (Chabchoub *et al.*, 2009; Wutz and Valencia, 2015; Vacca *et al.*, 2016). DNA methylation regulates expression of key regulator of inactivation of chromosome X – Xist (Beard, Li and Jaenisch, 1995; Nesterova *et al.*, 2008). Moreover, hypermethylation of CGI and gene promoters is a key event in X chromosome inactivation and correlates with transcriptional silencing (Sharp *et al.*, 2011; Cotton *et al.*, 2015). DNA methylation and DNMT activity are important for maintenance and stability of inactivation of chromosome X. Loss of DNA methylation induced by 5-azacitidine treatment or DNMT1 mutations leads to reactivation of inactivated X chromosome (Hansen *et al.*, 2000; Csankovszki, Nagy and Jaenisch, 2001). Given the profound effect of DNA methylation on phenotypic identity and cell functions, it is not surprising that aberrant DNA methylation and demethylation significantly affect cellular processes and can contribute to the development of pathologies including cancer. Cancer is characterised by uncontrolled proliferation and spread of cells that have managed to evade regulatory signals and gained a capacity for unstoppable growth, evade apoptosis and the immune response, and to disseminate to other organs. These dramatic changes in cell function are driven by genetic and epigenetic abnormalities that accumulate over time and ultimately lead to malignant transformation (Hanahan and Weinberg, 2011).

Epigenetic mechanisms are often disturbed in cancers and mutations in key epigenetic regulators including DNMT and TET enzymes and other chromatin-interacting proteins have been identified in various types of malignancies (Abdel-Wahab *et al.*, 2009; Roller *et al.*, 2013; Saghafeinia *et al.*, 2018; Han *et al.*, 2019). Impaired epigenetic regulation results in aberrant DNA methylation patterns characterised by global hypomethylation and local hypermethylation (Gama-sosa *et al.*, 1983; Greger *et al.*, 1989; Esteller *et al.*, 2001; Witte, Plass and Gerhauser, 2014). These methylation changes have a dual impact on the cancer genome. First, global methylation loss promotes reactivation of transposable elements, increases overall mutation burden and is associated with genomic instability (Rodriguez *et al.*, 2006; Ehrlich, 2009; Zhou *et al.*, 2018; Jang *et al.*, 2019). Second, local gain or loss of 5mC can drive oncogenic transcriptional programs and contribute to cancer development and progression (Figure 1-6). Loss of 5mC is frequently observed in some cancer types including Wilms' tumours (Ehrlich *et al.*, 2002), glioblastoma (Cadieux *et al.*, 2006) and liver cancer (Lin *et al.*, 2001). Decrease in DNA methylation can be also observed during cancer progression and development of metastasis (Lin *et al.*, 2001; Yegnasubramanian *et al.*, 2008; Vidal *et al.*, 2017). Moreover, both local hyper- and hypomethylation changes were reported in malignant cells (Keshet *et al.*, 2006; Saghafeinia *et al.*, 2018). For instance, cancer-specific, local DNA hypermethylation often occurs in transcriptional start sites (TSS), CGIs, enhancers, promoters and TF binding sites and may lead to silencing of genes involved in regulation of cell growth, apoptosis, cellular metabolism, signalling or cell adhesion. Similarly, local hypomethylation was proved to be associated with enhanced transcription of genes related to cell growth and proliferation, germ cell development or oncogenic functions (Saghafeinia *et al.*, 2018).

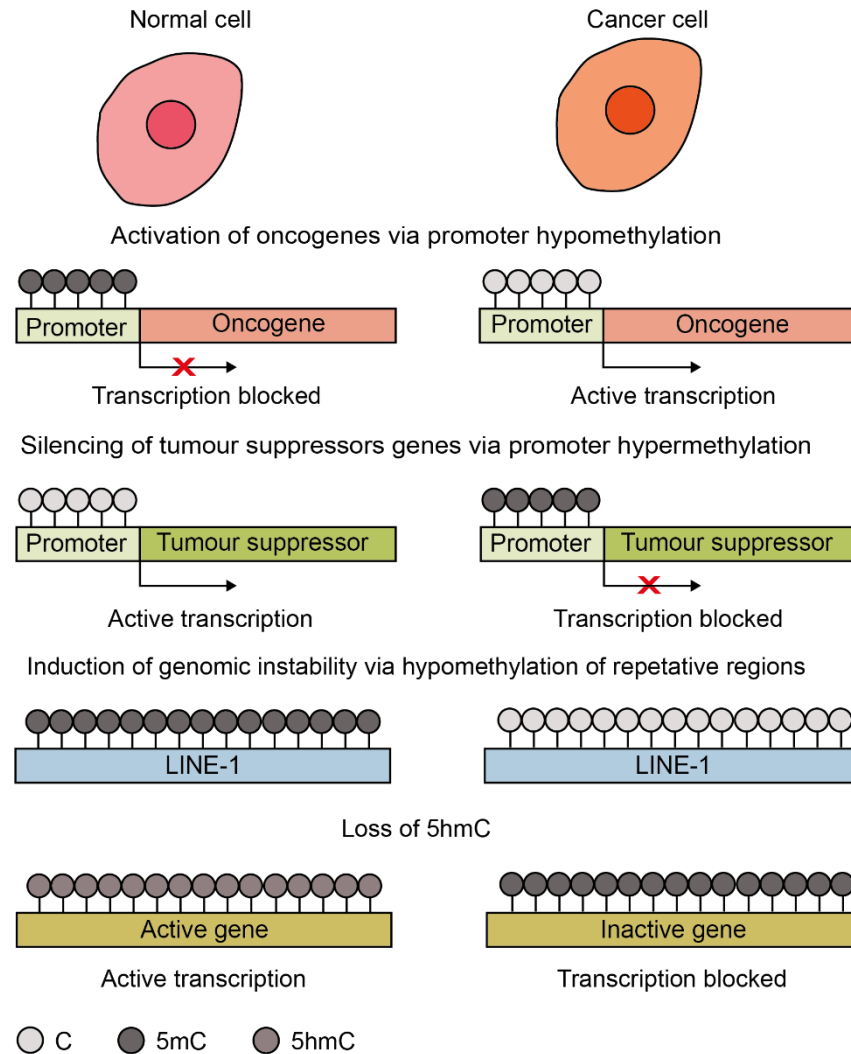


Figure 1-6. Summary of key epigenetic aberrations in cancer.

Another characteristic cancer-specific aberration is hypermethylation of CGIs and CGI shores that are unmethylated in normal cells (Keshet *et al.*, 2006; Skvortsova *et al.*, 2019). In addition, some tumours exhibit a unique epigenetic landscape with genome-wide hypermethylation in CGIs known as CpG island methylator phenotype (CIMP) (Hughes *et al.*, 2013). CIMP was described in many cancer types including colorectal cancer (Nosho *et al.*, 2009), leukaemia (Figueroa *et al.*, 2010; Fu *et al.*, 2014), glioma (Turcan *et al.*, 2012), breast cancer (Fang *et al.*, 2011), hepatocellular carcinoma (HCC) (Zhang *et al.*, 2007; Li *et al.*, 2019) and pancreatic ductal adenocarcinoma (PDAC) (Nicolle *et al.*, 2017; Lomberg *et al.*, 2018). Mutations in enzymes involved in deposition and

removal of DNA methylation are involved in development of CIMP (Nosho *et al.*, 2009; Figueroa *et al.*, 2010; Turcan *et al.*, 2012; Kai *et al.*, 2016). The most common genetic alterations observed in CIMP tumours are mutations in isocitrate dehydrogenase 1 gene (IDH1). The product of this gene is involved in generation of α KG which is necessary for TET activity. This mutant protein produces oncometabolite 2-hydroxyglutarate (2-HG) which inhibits TET enzymes and impairs DNA demethylation (Figueroa *et al.*, 2010; Turcan *et al.*, 2012; Hughes *et al.*, 2013). Moreover, direct mutations and transcriptional silencing of TET enzymes (Figueroa *et al.*, 2010; Kai *et al.*, 2016) as well as overexpression of DNMT proteins (Nosho *et al.*, 2009) were described as potential molecular changes underlying CIMP tumours. CIMP may produce different phenotypes in various cancer types, affect tumour invasiveness and influence patient prognosis (Fang *et al.*, 2011; Turcan *et al.*, 2012; Li *et al.*, 2019).

Together, cancer-specific DNA methylation changes can lead to establishment of new, transcriptional programs promoting a malignant phenotype (Keshet *et al.*, 2006; Teschendorff *et al.*, 2012; Saghafeinia *et al.*, 2018; Wang *et al.*, 2020).

Current studies suggest that aberrant changes in DNA methylation, especially local hypermethylation, occur early in cancer development and may significantly contribute to malignant transformation. Aberrant hypermethylation can be detected even in premalignant lesions (Sato *et al.*, 2002, 2008). In addition, increased DNA methylation variability has been observed more often in cervical swabs of patients progressing to neoplastic lesions years before morphological changes were detectable. Since DNA hypermethylation is often associated with gene silencing in cancer, altered methylation can potentially drive transcriptional programs that ultimately result in oncogenic transformation (Keshet *et al.*, 2006; Teschendorff *et al.*, 2010, 2012). Moreover, cancer-specific methylation often overlaps with alterations observed during normal aging,

suggesting that accumulation of DNA methylation changes are involved in age-related increased risk of cancer (Keshet *et al.*, 2006; Teschendorff *et al.*, 2010; Xie *et al.*, 2018; Zhou *et al.*, 2018).

Aberrant epigenetic patterns are proposed to be involved in cancer progression and among other factors determine metastatic potential of cancer cells (Fang *et al.*, 2011; Gkoutela *et al.*, 2019). For instance, breast cancers exhibiting CIMP phenotype are less likely to cause distant metastases and are associated with better survival than those with lower methylation levels, potentially due to epigenetic silencing of genes related to metastasis. This suggests that local hypermethylation can result in the attenuation of cell growth regulation and lead to a malignant phenotype at the initial stages of tumorigenesis, but simultaneously at later stages is associated with lower aggressiveness (Fang *et al.*, 2011). In addition, the recent study revealed that DNA methylation can be modified by the physical interactions between circulating tumours cells. Circulating tumour cells are often considered as an origin of distant metastasis and formation of cell clusters can induce methylation changes, including hypomethylation of pluripotency-associated and pro-proliferative genes and contribute to development of higher metastatic potential (Gkoutela *et al.*, 2019).

Aberrant DNA hydroxymethylation has also been observed in human cancers. Loss of hydroxymethylation has been observed in various malignancies and is associated with more advanced cancer stages, aggressive phenotype and worse prognosis (Ko *et al.*, 2010; Lian *et al.*, 2012; J. Liu *et al.*, 2019). On the other hand, Bhattacharyya *et al.* suggested that aberrant hydroxymethylation in PDAC, especially in enhancers, augments expression of oncogenes and contributes to malignant phenotype (Bhattacharyya *et al.*, 2017).

Finally, both aberrant methylation and hydroxymethylation patterns are important not only for cancer development and progression but also for maintenance of a malignant phenotype. Re-establishment of close-to-normal hydroxymethylation or methylation patterns leads to attenuation in the aggressive phenotype of cancer cells (Chen *et al.*, 2007; Lian *et al.*, 2012).

Due to the significant impact that abnormal methylation patterns have in cancer development and progression, DNA methylation is an interesting target for the development of new clinical approaches for cancer diagnosis, management and treatment. As DNA methylation changes are observed very early in cancerogenesis, methylation-based biomarkers offer promising approach for development cancer early detection methods.

1.2 DNA methylation for non-invasive cancer detection

1.2.1 Significance of early cancer diagnosis

Global incidence of cancer is increasing with more than 17M people diagnosed with cancer in 2018. Although the prognosis for cancer patients have improved over recent decades, mostly owing to development of novel therapeutic approaches, overall cancer mortality remains high: only in 2018 over 9M cancer-related deaths were reported worldwide (Bray *et al.*, 2018; Cancer Research UK, 2020).

High mortality rate can be at least partially attributed to the fact that cancer is often detected at an advanced stage (Figure 1-7), when tumour cells have already disseminated throughout the body. Numerous studies have demonstrated that cancers detected at earlier stage defined by smaller tumour size, lack of lymph node involvement, lack of distant metastases or other cancer type-specific scores are associated with significantly better prognosis and higher survival rate (Greten *et al.*, 2005; McPhail *et al.*, 2015; Hur *et al.*,

2016; Stark *et al.*, 2016; Shah *et al.*, 2017). On the other hand, only between 13.2-37.4% of common cancers including breast, lung, colorectal, prostate and ovarian tumours diagnosed in England in 2012 were at stage I.

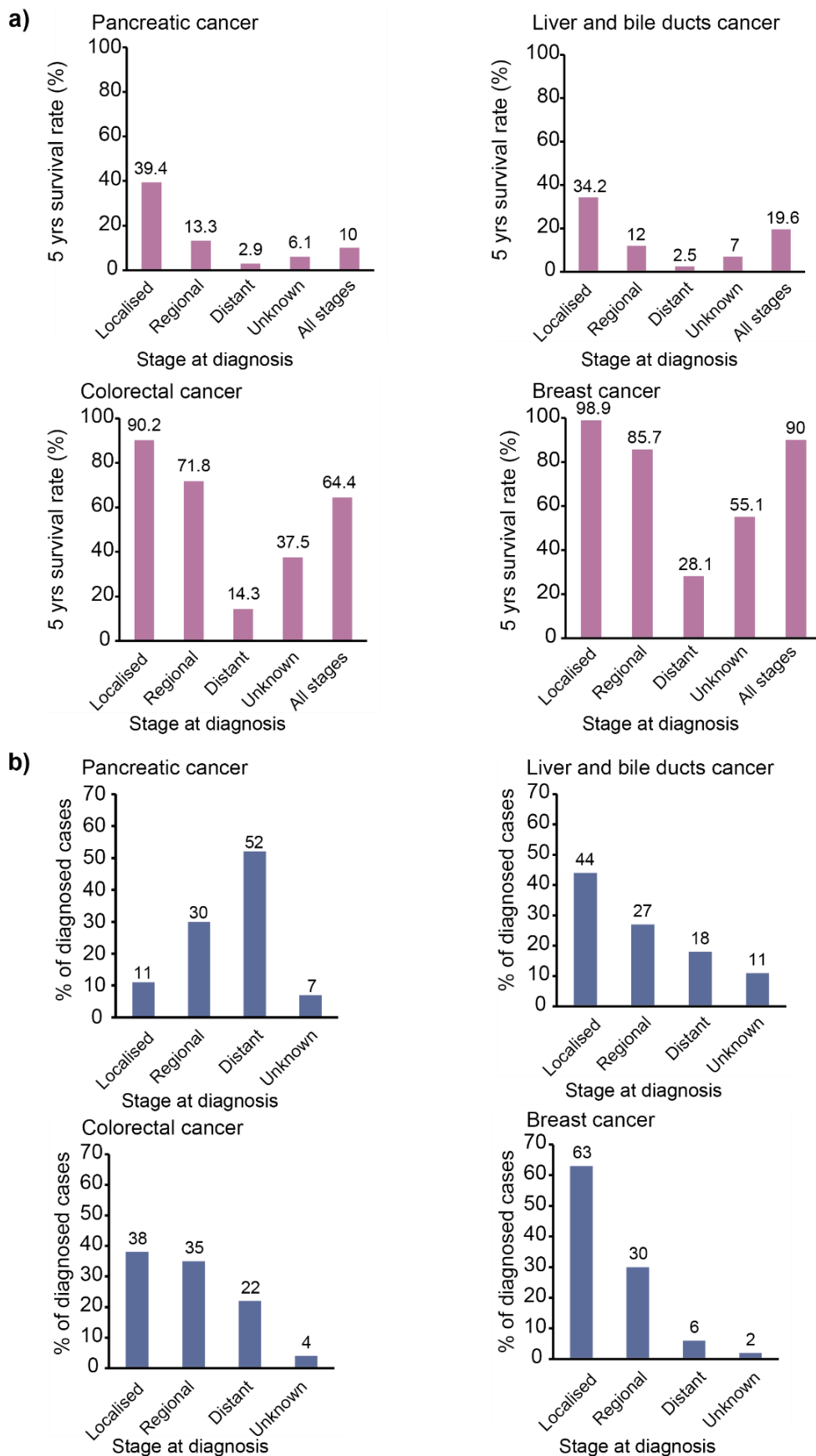


Figure 1-7. (a) 5-year survival rate depending on stage at diagnosis in four selected cancer types. (b) Stage at diagnosis for four selected cancer types. 5-year survival rate is better for patients with

localised disease. Cancer types for which screening programs or early detection methods are available (Breast cancer and colorectal cancer) are more often diagnosed at localised or regional stage and have better survival rate. Data sourced from SEER database (*Surveillance, Epidemiology, and End Results Program*, 2020).

Notably, stage at diagnosis differ significantly between cancer types with over 71% of breast cancer detected at stage I and II, possibly owing to existing screening programs. In contrast, nearly half newly reported cases of lung cancer were detected at stage IV and only around 20% at stage I (McPhail *et al.*, 2015). A similar trend was observed for invasive PDAC with approximately 50% of patients diagnosed with advanced tumours compared to 11% with localised disease and less than 2% diagnosed with curable precursor lesions, pancreatic intraepithelial neoplasia (PanIN) (Hur *et al.*, 2016; *Surveillance, Epidemiology, and End Results Program*, 2020).

Diagnosis at advanced stage significantly reduces survival rate and diminishes chances of successful treatment. For instance, median overall survival of patients diagnosed with PanIN is around 120 months but it drops down to 11 months for patients diagnosed with invasive pancreatic tumours bigger than 2 cm (Hur *et al.*, 2016). Similarly, overall survival for patients diagnosed with localised hepatocellular carcinoma between 2008-2013 was 31 months compared to only 5 months for patients diagnosed with metastatic disease (Shah *et al.*, 2017).

The close relationship between stage at diagnosis and patient prognosis reflects an urgent need for development of novel diagnostic methods for more sensitive early detection of cancer.

1.2.2 Liquid biopsy for cancer diagnosis and detection

Unfortunately, sensitive diagnostic tests able to detect cancer early, ideally at asymptomatic stage, are yet to be developed for most cancer types. Current methods for detection and diagnosis of cancer are often not suitable for asymptomatic population

screening. For instance, imaging-based technologies such as Magnetic Resonance Imaging (MRI) or Computed Tomography (CT), widely used for detection of malignant lesions or disease relapse, are not efficient in detection of very small, preclinical tumours due to limited sensitivity and relatively high costs. Similarly, ultrasound which is a useful and relatively cheap diagnostic tool may not be sensitive enough for detection of small lesions. For instance, ultrasound has an overall sensitivity of 94% to detect HCC at any stage, however, only 64% sensitivity to detect early stage tumours (Pascual *et al.*, 2019). Moreover, ability of ultrasound to detect small cancers in deeply located internal organs such as the pancreas is rather limited. Additionally, some imaging procedures are considered invasive as they require exogenous contrast to improve signal-to-background ratio or involve radiation exposure (Fass, 2008; Frangioni, 2008; Lim and Singal, 2019). Therefore, novel, minimally invasive approaches suitable for early diagnosis and monitoring of cancer are needed.

Liquid biopsy is considered as a promising direction in development of new diagnostic tools. It is based on detection of cancer biomarkers in easily accessible body fluids such as blood, urine, saliva, cerebrospinal fluid or stool (Imperiale *et al.*, 2014; De Mattos-Arruda *et al.*, 2015; Yu *et al.*, 2016; Wan *et al.*, 2017; Cohen *et al.*, 2018; Nuzzo *et al.*, 2020). Liquid biopsy can involve analysis of a wide range of analytes including proteins, cell-free DNA (cfDNA), circulating tumour cells (CTCs), circulating tumour RNA (cfRNA) and exosomes (Riethdorf *et al.*, 2007; Siravegna *et al.*, 2017; Wan *et al.*, 2017; Cohen *et al.*, 2018; Heitzer *et al.*, 2018; D. S. K. Liu *et al.*, 2019). Moreover, liquid biopsy can be applied at different stages of the clinical process from initial cancer detection and patient stratification to treatment monitoring and detection of relapse (Figure 1-8) (Riethdorf *et al.*, 2007; Olsson *et al.*, 2015; Cohen, Javed, Thoburn, Wong, Tie, Gibbs,

et al., 2017; Phallen *et al.*, 2017; Siravegna *et al.*, 2017; Cohen *et al.*, 2018; Parikh *et al.*, 2019).

STAGE	GOAL
Asymptomatic person	Screening Cancer detection
Confirmed cancer	Treatment selection Prognosis
Cancer treatment	Response monitoring Detection of resistance Detection of MRD
Posttreatment monitoring	Detection of metastasis Qualification for adjuvant therapy

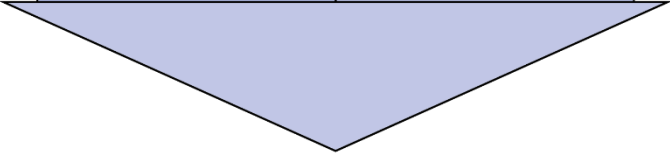


Figure 1-8. Potential applications of liquid biopsy at different stages of cancer management.

Although liquid biopsy will not fully replace standard tissue biopsy in clinical practice, it could overcome some limitations of traditional approaches and thus become an invaluable complementary method. The main limitation of tissue biopsy is that it can be performed only with detectable tumour, so by definition it lacks early diagnosis potential. In contrast, liquid biopsy relies on analysis of body fluids and has potential for initial detection of cancer or when the location of original tumour is unknown. Another drawback of tissue biopsy is its invasiveness and associated risk of side-effects. Therefore, this procedure cannot be repeated multiple times. By contrast, liquid biopsy is minimally invasive and could be particularly useful when multiple samples are required to be taken in a short period of time e.g. during real-time treatment monitoring. Finally, standard tissue biopsy usually picks only part of the tumour and may fail to capture tumour heterogeneity with negative effect on diagnosis or treatment decisions. Liquid biopsy has the potential to overcome this limitation: some studies have already

demonstrated that it outperforms standard tissue biopsy to capture inter- and intratumor heterogeneity and provide clinically useful information (De Mattos-Arruda *et al.*, 2015; Siravegna *et al.*, 2017; Heitzer *et al.*, 2018; Mannelli, 2019; Parikh *et al.*, 2019). These examples illustrate the large potential utility of liquid biopsy as a diagnostic tool.

1.2.3 Cell-free DNA as cancer biomarker

CfDNA is one of the most extensively studied candidates for development of liquid biopsy assays. Presence of circulating nucleic acids in the blood was first reported in 1948 by Mandel and Metais. Almost 50 years later, two studies reported presence of cancer-specific RAS mutations in cfDNA from cancer patients (Sorenson *et al.*, 1994; Vasioukhin *et al.*, 1994; Thierry *et al.*, 2016; van der Pol and Mouliere, 2019). Since then, the idea of using cfDNA as a cancer biomarker has been explored by researchers around the world.

CfDNA are short DNA fragments found in blood or other body fluids such as urine, saliva, stool or cerebrospinal fluid. Cells release cfDNA fragments to the bloodstream through various mechanisms including apoptosis, necrosis or secretion. CfDNA is then rapidly removed from the blood with a half-life time around 2 h (Diehl *et al.*, 2008; De Mattos-Arruda *et al.*, 2015; Thierry *et al.*, 2016; Siravegna *et al.*, 2017; Wan *et al.*, 2017; van der Pol and Mouliere, 2019).

Plasma cfDNA exhibits characteristic fragmentation profile with a major peak at 166-168 bp, which corresponds to length of DNA wrapped around nucleosome and linker DNA. However, both shorter and longer cfDNA fragments can be found and their length depend mostly on release mechanism, cell-of-origin or chromatin state in original cell (Jiang *et al.*, 2015; Thierry *et al.*, 2016; Underhill *et al.*, 2016; Mouliere *et al.*, 2018; Han *et al.*, 2020). Overall level of cfDNA is low in healthy individuals but increases in

pathological states such as cancer, sepsis, tissue injury or inflammation (Wan *et al.*, 2017). The majority of cfDNA comes from blood cells, however, in the presence of cancer, tumour cells release circulating tumour DNA (ctDNA) that contributes to total cfDNA pool. Detection of cancer-specific markers including somatic mutations, chromosome rearrangement, methylation patterns, fragmentation profiles and other features in cfDNA is considered a promising non-invasive approach for cancer diagnosis and detection (Diehl *et al.*, 2008; S. Guo *et al.*, 2017; Wan *et al.*, 2017; Moss *et al.*, 2018; van der Pol and Mouliere, 2019).

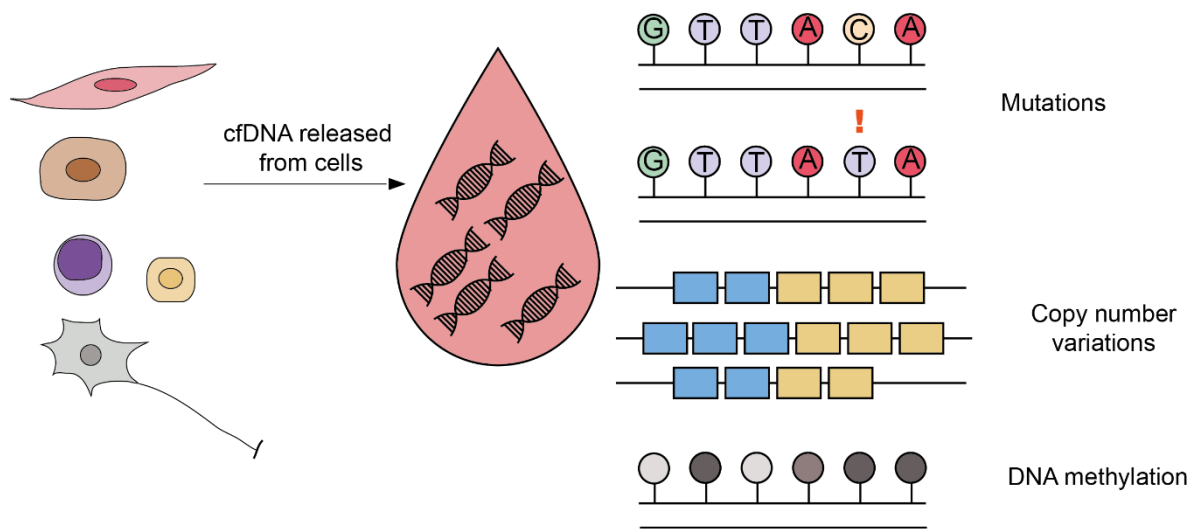


Figure 1-9. CfDNA summary. CfDNA fragments are released to the circulation from various cells in the body. Mutations, CNVs and DNA methylation found in cfDNA reflect genetic and epigenetic status of cell of origin.

Detection of cancer-specific mutations or other genetic changes in cfDNA is a common direction in development of the liquid biopsy assays. Some of those methods have proved their utility for cancer management including patient stratification, treatment monitoring, detection of minimal residual disease (MRD) or early detection of distant metastasis (Olsson *et al.*, 2015; Parikh *et al.*, 2019; Rothwell *et al.*, 2019; Zviran, Rafael C. Schulman, *et al.*, 2020). For instance, cobas EGFR Mutation test v2 that detects Epidermal Growth Factor Receptor (EGFR) mutations in plasma ctDNA has been

approved by US Food and Drug Administration (FDA) to be used in non-small-cell lung carcinoma (NSCLC) to identify patients for treatment with tyrosine kinase inhibitors (TKIs) (Torres *et al.*, 2020). Furthermore, post-surgical or post-treatment monitoring of ctDNA level can help to identify patients with MRD or at high risk of recurrence. In fact, increased level of ctDNA in patients with MRD could be observed significantly earlier than detection of relapse by imaging methods. For instance, the appearance of ctDNA can precede clinically detectable metastases even up to 3 years earlier in breast cancer patients (Olsson *et al.*, 2015; Zviran, Rafael C. Schulman, *et al.*, 2020). Finally, cfDNA analysis helps to detect mutations related to development of treatment-resistance and in the future could be used for real-time treatment monitoring and support treatment decisions, to increase chances for cancer to be completely eradicated (De Mattos-Arruda *et al.*, 2015; Chabon *et al.*, 2016; Parikh *et al.*, 2019).

Despite successful application of cfDNA for monitoring purposes, the identification cancer-specific mutations in cfDNA for early cancer detection is more challenging and not widely used in clinical practice. CtDNA constitutes only a small fraction of circulating DNA and tumour-derived mutated fragments are diluted in the background of wild-type cfDNA. Although ctDNA levels differ between patients and depend on cancer type, tumour size and disease stage, tumour contribution to cfDNA is usually very low at the initial stages of cancerogenesis (Diehl *et al.*, 2005; Thierry *et al.*, 2016; Phallen *et al.*, 2017). According to the review by Kustanovich *et al.*, ctDNA fraction in cfDNA of cancer patients varies between 0.1-89% (Kustanovich *et al.*, 2019). However, tumour contribution in plasma can be far below 0.1% in non-advanced disease. For example, the average mutant fraction in early stage colorectal cancer is as low as 0.01% (ranging from 0.01%-0.12%), compared to 1.9-27% in advanced, metastatic disease (Diehl *et al.*, 2005). Moreover, not all patients have detectable levels of ctDNA. Bettegowda *et al.* have shown

that although ctDNA is detectable in 82% of patients with metastatic cancer, it is detected in only 47% of patients with localised cancer using digital PCR-based approaches (Bettegowda *et al.*, 2014). Notably, most currently existing methods can detect ctDNA with low detection limit ranging from 0.0001-3%. PCR-based methods are among the most sensitive but can detect only previously identified candidate mutations (Siravegna *et al.*, 2017). This means that ctDNA levels in many patients are below the sensitivity limit and will contribute to increased number false negative results (Bettegowda *et al.*, 2014; Zviran, Rafael C. Schulman, *et al.*, 2020). A second disadvantage of mutation-based ctDNA tests is the fact that mutations cannot inform about tissue of origin, which would be problematic in pan-cancer diagnostic tests. Finally, as recent reports have proved accumulation of somatic mutations in hematopoietic cells, mutations derived from clonal haematopoiesis can be mistaken for cancer-derived and lead to false positives outcomes and unnecessary medical interventions (Hu *et al.*, 2018).

A common approach to improve sensitivity of mutation-based ctDNA assays is combining multiple genetic biomarkers into single assay. However, to date even assays based on gene panels have failed to achieve high sensitivity in early stage cancers. For example, the gene panel developed by Phallen *et al.* detected 62% of stage I and II cancers compared to 77% of stage III and IV (Phallen *et al.*, 2017). In turn, Cohen *et al.* developed multianalyte liquid biopsy test called CancerSEEK combining detection of 1933 genetic changes including point mutations, insertions and deletions in 16 cancer genes in ctDNA with assessment of 8 protein serum biomarkers (Cohen *et al.*, 2018). Assay evaluation in a cohort of over 1000 cancer patients diagnosed with the eight most common cancer types and over 800 healthy controls achieved 70% sensitivity at specificity over 99%. On the other hand, CancerSEEK sensitivity for stage I tumours was only 47%, which means that over half of early stage tumours were left undetected by this method.

Another interesting direction that could increase sensitivity of genomic-based liquid biopsy tests is using cfDNA fragmentation profiles for cancer prediction. Recent research has suggested that tumour-derived cfDNA fragments are shorter compared to the major 167 bp fraction and can be possibly used as a marker for cancer (Jiang *et al.*, 2015; Underhill *et al.*, 2016; Mouliere *et al.*, 2018). Moreover, selecting for shorter fragments could help to enrich for cancer-specific genetic changes. As fragmentation profiles can be retrieved even from shallow sequencing, fragmentation analysis offers a cost-effective alternative to detection of somatic mutations. Mouliere *et al.* have demonstrated that selection of cfDNA fragments 90-150 bp helped to enrich for tumour-specific circulating DNA and increased detection of CNVs and cancer mutations. Combining fragmentation profiles with genetic features exhibited improved disease prediction compared to prediction based only on genetic changes. The proposed assay achieved 94% sensitivity and specificity 95% for cancer types characterised by high levels of ctDNA and 64 % sensitivity for cancers with low ctDNA level (Mouliere *et al.*, 2018). Another seminal study in the field published by Cristiano *et al.* utilised fragmentation profiles as the only predictor for cancer and achieved cancer detection sensitivity of 57-99% at specificity 98% suggesting that cfDNA fragmentation itself can serve as cancer biomarker. This assay performed particularly well for lung and ovarian cancer with sensitivity of 100% and 89%, respectively. Notably this approach detected more early stage cancers compared to above-mentioned mutation based-assay with 68% sensitivity at stage I (Cristiano *et al.*, 2019).

In summary, although detection of cancer-specific genetic mutations in cfDNA is useful for cancer monitoring and management, it is challenging for cancer detection, especially at early stages of disease. Therefore, other features of cfDNA including epigenetic marks are being considered as novel candidates for biomarkers search in liquid biopsy field.

1.2.4 cfDNA methylation signatures in liquid biopsy

Methylation analysis of cfDNA is considered as a promising direction in liquid biopsy assays. Cancer-specific methylation changes at global and local level have been thoroughly studied in human malignancies (van der Pol and Mouliere, 2019). Tumour-specific methylation and hydroxymethylation changes are detectable in cfDNA of cancer patients justifying further development of epigenetic-based liquid biopsy assays (Shen *et al.*, 2018). Analysis of DNA methylation in cfDNA offers advantages compared to detection of genomic alterations (Klein *et al.*, 2018). In general methylation changes occur early in cancerogenesis process, even before accumulation of genetic changes. Changes in methylation signatures therefore have potential for cancer detection earlier than somatic mutations (van der Pol and Mouliere, 2019). Moreover, DNA methylation signatures are tissue-specific, which could be particularly useful in diagnostic tests where cancer origin has to be identified (Sun *et al.*, 2015; S. Guo *et al.*, 2017; Moss *et al.*, 2018). Additionally, tissue-specific methylation signatures can serve as stand-alone cancer biomarkers. During cancerogenesis level of tissue-specific DNA in the blood may increase due to tissue remodelling in the affected organ and DNA release from cancer cells. For instance, increase in tissue-specific cfDNA level assessed based on single tissue-specific methylation biomarker could be detected at stage I and II of PDAC while KRAS mutations could be only found at stage IV (Lehmann-Werman *et al.*, 2016). Finally, as methylation changes can be triggered by specific stimuli, it is possible that ongoing cancerogenesis process affects the epigenetic landscape of immune cells and that these changes could be detected in cfDNA (Koestler *et al.*, 2012; Y. Zhang *et al.*, 2018). In fact, presence of cancer-induced transcriptional profiles in platelets gives rise to tumour-educated platelets (Best *et al.*, 2015). In the presence of cancer, similar events could occur at epigenetic level in different cell types. As most of cfDNA origins from

immune and blood cells, detection of cancer-induced methylation signatures derived from these cells could help to detect early stage tumours even if ctDNA level is low.

Despite the great interest in epigenetics-based liquid biopsy assays, to date only two diagnostic tests including methylation analysis have been approved by FDA, both for detection of colon cancer (Koch *et al.*, 2018). Epi proColon is PCR-based assay detecting methylation status of SEPT9 promoter in plasma cfDNA. Epi proColon performs worse than traditional methods for detection of colorectal cancer (CRC) with 70% sensitivity and 80% specificity, however, it is still recommended as an alternative to lowly acceptable diagnostic procedures such as colonoscopy or occult blood detection in faecal samples (Shirley, 2020). The second approved test, Cologuard, assesses methylation status of two genes NDRG4 and BMP3 as well as 7 somatic mutations in stool cfDNA. Cologuard detects CRC with sensitivity of 92.3% and specificity around 90%. Notably, sensitivity of this assay outperforms faecal immunochemical test (FIT), however, specificity is slightly lower (Imperiale *et al.*, 2014; Pickhardt, 2016).

Numerous other methylation-based cfDNA assays are now under development. The Circulating Cell-Free DNA Genome Atlas (CCGA) study designed and validated a diagnostic test based on target methylation analysis of over 100 000 genomic regions covering more than 1 000 000 individual CpG sites. The study involved over 10 000 participants including 50 cancer types and healthy controls. CCGA confirmed good ability of methylation-based assays to predict tissue-of-origin with 93% accuracy. On the other hand, sensitivity of the panel depended on disease stage and ranged from 18% (13-25%) at stage I to 93% (86-96%) at stage IV, with specificity fixed at over 99%. Although sensitivity at stage I for all cancer types was low, this assay had promising results for some stage I cancers including PDAC and head and neck tumours with sensitivity 55-70%. This study confirmed feasibility of cfDNA methylation to be used for cancer

detection at high specificity, however, sensitivity to detect non-advanced disease was disappointingly low and requires further improvements (M. C. Liu *et al.*, 2020).

Three other studies used cell-free Methylated DNA Immunoprecipitation and High-throughput Sequencing (cfMeDIP) approach to compare cfDNA methylation profiles in cancer and healthy (Li, Li and Xu, 2018; Shen *et al.*, 2018; Nassiri *et al.*, 2020; Nuzzo *et al.*, 2020). The advantage of this approach is the bisulfite-free protocol that allows the decrease cfDNA input to 1-10 ng. CfMeDIP achieved promising AUC values 0.918-0.99 even for challenging, low ctDNA cancers such as renal cell carcinoma (RCC) or intracranial tumours. On the other hand, all three studies have important limitations. First, tumour classifications were based on direct comparison of methylation profiles in cancer and healthy on multiple training and validation sets and final methylation-based biomarker panels were not proposed. Moreover, a significant difference (over 15 years) in medium age of control and cancer group in RCC study could contribute to observed, high AUC. Since methylation patterns change during aging, it is impossible to determine whether sample classification was driven by cancer or age difference (Nuzzo *et al.*, 2020). In addition to 5mC, cancer-specific cfDNA hydroxymethylation changes can be used for detection purposes. Global loss in cfDNA 5hmC was observed in lung cancer, and locus-specific loss and gain of 5hmC was reported in PDAC and HCC. These cancer-specific 5hmC patterns could accurately predict cancer type and stage (Song *et al.*, 2017).

Unfortunately, despite initial promising results, most methylation biomarkers described in the literature have not been translated into clinically applicable diagnostic tools. As reviewed by Koch *et al.*, there were nearly 15 000 research papers proposing around 1800 methylation-based biomarkers in PubMed in January 2018. However, only 14 methylation-based tests are commercially available, with 9 included into clinical guidelines and only 2 FDA-approved (Koch *et al.*, 2018). This disappointing bench-to-

bed translation rate does not necessarily mean that development of highly sensitive and specific methylation-based diagnostic tests is not feasible. Possibly, the correct biomarkers have yet to be identified or multianalyte tests combining cfDNA methylation profiles with other genomic features have to be developed. Unfortunately, most current studies tend to focus on methylation-only biomarkers or restrict candidate biomarkers to differentially methylated regions identified based on tumour data. When using this approach, potentially informative, cancer-specific methylation changes in cfDNA derived from other cell populations e.g. immune cells are largely overlooked. Comprehensive whole-genome methylome profiling of cfDNA in cancer and healthy could help to identify the cancer specific methylation signatures in cfDNA in unbiased way. However, such studies are limited due to technological challenges.

CfDNA amount is low and usually does not exceed 10 ng in 1 mL of plasma. Therefore, it is problematic to analyse by bisulfite-sequencing (BS-seq) – the gold standard for whole-genome methylome analysis, due to harsh chemical reaction and excessive DNA damage involved in BS-seq protocol (Worm Ørntoft *et al.*, 2017). CfMeDIP method offers mild, non-damaging protocol, but as an enrichment-based method it misses base-resolution information about methylation and hydroxymethylation. Additionally, current methods for methylation mapping do not enable for multidimensional analysis of cfDNA including not only methylation but also other features such as fragment length, genetic mutations and CNVs in a single assay. Therefore, further technological improvements are required to enable comprehensive analysis in cfDNA necessary for identification of novel cfDNA cancer biomarkers.

1.2.5 Diagnosis and management of pancreatic cancer

Pancreatic cancer mostly refers to pancreatic ductal adenocarcinoma (PDAC) which is the most common type of cancer that develops in pancreas. PDAC is one of the deadliest malignant disease and was responsible for around 4.5% of cancer-related deaths in 2018. Overall prognosis for PDAC are poor, with only 9-13% of patients surviving 5 years and overall median survival of 11 months. Despite recent progress in cancer treatment and diagnosis, prognosis for PDAC patients has not improved significantly. PDAC survival rate is negatively affected by advanced disease stage as defined based on tumour size and lymph node involvement. At the same time most of PDAC patients are diagnosed at advanced stage with only 10-20% of patients having resectable tumours. Notably, survival rate for people who underwent surgery is still low and only around 22% and 12.1% of patients survive for 5 and 10 years, respectively. Long-term survivors often display a lack of lymph node involvement, absence of perineural invasion and low tumour grade. However, the low proportion of PDAC patients that survived for more than 5 years after diagnosis means that with current diagnostic techniques even resectable tumours are detected too late when curative treatment is already not possible (Stark *et al.*, 2016; Neoptolemos *et al.*, 2018; Rawla, Sunkara and Gaduputi, 2019).

Unfortunately, methods for early detection of pancreatic cancer are not available. Anatomical location of the pancreas deep inside abdominal cavity and lack of specific symptoms at non-advanced stage impede early PDAC detection (Evans *et al.*, 2014; Neoptolemos *et al.*, 2018). At later, symptomatic stage, imaging techniques including CT, MRI or PET are often used, however, they are not feasible for general screening and are not sensitive enough to detect small *in situ* lesions (Hur *et al.*, 2016; Ngamruengphong and Canto, 2016). Serum Carbohydrate antigen 19-9 (CA19-9) is sometimes considered as PDAC biomarker, but its sensitivity and specificity are too low for general screening

and are estimated as 47-79% and 82-88%, respectively. Moreover, CA19-9 is often elevated in non-malignant conditions including inflammatory disease and 5-10% people are unable to produce this antigen at all (Goonetilleke and Siriwardena, 2007; Cao *et al.*, 2016; Ngamruengphong and Canto, 2016).

Some studies have shown that combination of CA19-9 and other markers offers improved performance compared to CA19-9 itself. For instance, simultaneous assessment of CA19-9 and thrombospondin-2 enabled to detect PDAC with 87% sensitivity and 98% specificity and more reliable discrimination between PDAC and pancreatitis. However, this panel has to be further validated and is not widely used in the clinical settings (Kim *et al.*, 2017).

Molecular changes contributing to development of PDAC include activation of pro-proliferation pathways, suppression of pro-apoptotic pathways, impaired differentiation and immune suppression (Greten *et al.*, 2002; Gore *et al.*, 2014; Bailey *et al.*, 2016; Neoptolemos *et al.*, 2018). For instance, aberrant Tumour Growth Factor beta (TGF- β) signalling in the background of inactive retinoblastoma 1 (RB) gene leads to upregulation of Wnt7b in SMAD4-dependant manner and activates other oncogenic pathways such as MAPK/ERK or Src. This aberrant signalling contributes to enhanced proliferation and invasiveness of pancreatic cells (Gore *et al.*, 2014). Some pancreatic tumours also exhibit significant changes in immune signalling including antigen presentation, Toll-like receptor signalling, B cell signalling as well as upregulation of CTLA4 and PD1 related to tumour immune suppression (Bailey *et al.*, 2016). Activation of SMAD3 and NF- κ B signalling accompanied by overexpression of Bcl-x_l promotes resistance to apoptosis and development of invasive tumours (Greten *et al.*, 2002).

The genetic background of PDAC is complex. Bailey *et al.* performed genomic analysis of over 380 pancreatic tumours and identified more than 20 000 mutations and genomic

rearrangements. However, only 32 genes were recurrently mutated in the final cohort of 456 tumours. These recurrent genetic changes included oncogenic KRAS mutations detected in over 92% of tumours, inactivating mutations in G1/S checkpoint genes: TP53, CDKN2A and TP53BP2 in 78% of tumours and mutations of TGF- β signalling genes in 47% of tumours, making them attractive markers for ctDNA-based PDAC detection (Bailey *et al.*, 2016). Unfortunately, KRAS mutations in ctDNA are detected in 22-50% of patients, mostly at advanced stage limiting its application as early detection biomarker (Lehmann-Werman *et al.*, 2016; Cohen, Javed, Thoburn, Wong, Tie, Lennon, *et al.*, 2017; Shinjo *et al.*, 2020). Low sensitivity of KRAS detection can be partially caused by the fact that pancreatic tumours shed limited amount of DNA into circulation (Mouliere *et al.*, 2018). Better sensitivity was achieved from CancerSEEK panel combining genetic mutations and protein biomarkers. This assay identified 64% of resectable PDAC cases at 98% specificity (Cohen, Javed, Thoburn, Wong, Tie, Gibbs, *et al.*, 2017).

Epigenetic changes are important for PDAC development as well. DNA methylation changes can occur early even at the level of pre-neoplastic lesions and in some cases can be driven by pancreatitis which is one of the risk factors for PDAC (Sato *et al.*, 2002, 2008; Natale *et al.*, 2019; Bararia *et al.*, 2020). Progressive aberrant hypermethylation of CGIs in genes involved in cell cycle regulation, adhesion, apoptosis, neuronal function and tight junctions formation occur during progression from PanIN to PDAC (Sato *et al.*, 2002, 2008). Locus-specific gain of 5hmC can lead to activation of oncogene expression (Bhattacharyya *et al.*, 2017) while changes in DNA methylation, especially in promoters and enhancer, together with histone marks influence gene expression and can activate genes with oncogenic, pro-proliferative potential (e.g. ERBB, FOS and CREB) and silence proapoptotic and tumour suppressor (e.g. WT1, CDKN2A, STK11, IRF5) genes. Aberrant methylation in PDAC disturbs key cell signalling pathways including TGF- β

signalling, WNT/NOTCH signalling, axon guidance, cell cycle regulation, apoptosis and pluripotency (Vincent *et al.*, 2011; Nones *et al.*, 2014; Tang *et al.*, 2015; Lomberk *et al.*, 2018).

For instance, CDKN2A gene involved in cell cycle regulation is often mutated in PDAC. However, in some patients it can be inactivated by promoter hypermethylation and hypermethylation of CDKN2A is associated with poor prognosis (Tang *et al.*, 2015). Moreover, according to study by Nones *et al.* genes involved in stellate cell activation including EGFR and MET are hypomethylated in PDAC. In turn, key genes in SLIT-ROBO pathway involved in axon guidance and regulation of WNT and HGF/MET signalling are inactivated by hypermethylation (Nones *et al.*, 2014). Hypomethylation of reprogramming regulators KLF4 and SOX9 (Lomberk *et al.*, 2018) and hypermethylation of other pluripotency genes e.g. BMP3/6 or FOXD3 (Vincent *et al.*, 2011) were both observed in pancreatic cancer suggesting aberrant pluripotency pathways.

Epigenetic mechanisms also play a role in development of different PDAC subtypes, and epigenetic modulators including DNMT1 and histone remodellers are often upregulated in PDAC, further contributing to aberrant methylation changes observed in this malignancy (Lomberk *et al.*, 2018). Moreover, abnormal methylation signatures in PDAC may be associated with metastatic potential (McDonald *et al.*, 2017).

Some studies already confirmed that methylation in cfDNA can be useful biomarker for PDAC. Pancreas-specific methylation biomarkers can be found in blood of PDAC patients at earlier stage when KRAS mutations are not yet detectable (Lehmann-Werman *et al.*, 2016). Similarly, target methylation assay in DNA from pancreatic juice shows better sensitivity compared to detection of KRAS mutations (Majumder *et al.*, 2019). The CCGA study demonstrated 68-100% sensitivity to detect PDAC at 98% specificity with targeted methylation analysis confirming feasibility of early PDAC detection based on

methylation signatures (M. C. Liu *et al.*, 2020). Another methylation-based cfDNA study using cfMeDIP method has shown that methylation signatures can be used for detection of PDAC with AUC = 0.915 (Shen *et al.*, 2018). In contrast, a methylation panel that included 5 aberrantly methylated genes, defined based on tumour data combined with analysis of KRAS mutations, showed 66% sensitivity and 86% specificity (Shinjo *et al.*, 2020). This suggests that small gene panels limited to the biomarkers selected by tumour-informed approaches perform worse than assays which include features identified directly from cfDNA or those covering multiple genomic loci. In fact, recent studies have shown that increased sequencing breadth and whole genome coverage, as well as targeted assays analysing numerous loci, can facilitate ctDNA detection at low tumour burden (M. C. Liu *et al.*, 2020; Wan *et al.*, 2020; Zviran, Rafael C Schulman, *et al.*, 2020). Moreover, according to some results, cancer prediction based on the features identified by tumour-naïve approaches have better prediction accuracy than panels limited to tumour-specific aberrations (Chen *et al.*, 2020; Guler *et al.*, 2020).

Taken together, cfDNA methylation biomarkers emerged as promising alternative for PDAC diagnosis, often outperforming mutations or protein-based biomarker. On the other hand, further developments are required for methylation-based assays to meet clinical utility criteria.

1.2.6 Diagnosis and management of hepatocellular carcinoma

Hepatocellular carcinoma (HCC) is one of the most common cancer and third cause of cancer-related deaths worldwide. HCC risk factors include HBV and HCV infections, liver cirrhosis and high alcohol intake. HCC often affects males and around three quarters of HCC patients are male. Overall survival rate for HCC patients is low with survival of 11-14 months, and is usually better for early stage disease (Greten *et al.*, 2005; Shah *et*

al., 2017). Specifically, survival at early stage localised disease in 2008-2013 was 31 months compared with 5 months for metastatic disease. Notably, there was significant survival improvement of 8 months for patients at localised stage in 2001-2007, partially associated with novel therapies, which was not observed for metastatic patients. This indicates importance of early HCC diagnosis for clinical benefit (Shah *et al.*, 2017).

Currently screening for HCC in whole population is not available, however, high risk groups including people with liver cirrhosis and chronic liver disease are often under surveillance programs to detect developing cancer early. These usually include imaging methods such as ultrasound or CT with or without monitoring of alpha-fetoprotein (AFP) serum level. However, it is important to highlight that sensitivities of these approaches are still limited. Ultrasound is able to detect 94% of HCC at all stages however only 63% at early stage with 94% specificity (Pascual *et al.*, 2019). Sensitivity of CT for HCC detection is slightly higher, however, it is also more expensive and overall detection improvement is not statistically significant (Pocha *et al.*, 2013). The increased level of blood biomarker, AFP, can detect HCC with 60% sensitivity and 90.6% specificity, however, it may contribute to false results in patients with chronic liver diseases. The combination of AFP and imaging techniques did not improve sensitivity and specificity. Moreover, although combined screening facilitated earlier cancer detection, survival benefit was rather limited (Trevisani *et al.*, 2001; Kansagara *et al.*, 2014; Pascual *et al.*, 2019).

Numerous genetic and epigenetic abnormalities have been observed in development and progression of HCC. Chronic inflammation triggered by viral infections, lifestyle and diet promotes accumulation of genetic events that drive tumorigenesis. HCC tumours exhibit heterogenous mutational profiles with mutations occurring in tumour suppressor genes, chromatin remodellers, genes involved in antioxidant response and key signalling

pathways. Amplification of oncogenes and deletions of tumour suppressor genes have been described as well. The most common mutations are TERT promoter mutations associated with cell immortalisation as well as inactivating mutations in TP53. All observed genetic changes result in dysregulation of key signalling pathways promoting proliferation and survival and preventing growth regulation and cell death (Whittaker, Marais and Zhu, 2010; Schulze, Nault and Villanueva, 2016; Ally *et al.*, 2017). HCC cells exhibit heterogenous methylation, CNV and transcriptional profiles and genome-wide hypomethylation (Hou *et al.*, 2016).

Epigenetic changes are common in HCC and include genome-wide hypomethylation, hypermethylation of tumour suppressor genes and hypomethylation of LINE-1 elements (Lin *et al.*, 2001; Yang *et al.*, 2003; Gao *et al.*, 2014). Epigenetic silencing of CDKN2A gene often accompanying TERT mutations is frequently observed in HCC (Ally *et al.*, 2017). A decreased level of 5hmC and other oxidised cytosine variants has been observed in HCC, potentially caused by decrease in TET2 expression and dysregulation of other epigenetic enzymes (J. Liu *et al.*, 2019).

Attempts to use liquid biopsy approaches for detection of HCC have been made, as cancer-specific epigenetic and genetic changes are observed in cfDNA of HCC patients (Kaseb *et al.*, 2019). Shortening of cfDNA fragments was described in HCC, and the observed correlation between proportion of short fragments and tumour fractions suggested that those short fragments are likely of tumour origin (Jiang *et al.*, 2015).

CancerSEEK panel shows excellent sensitivity for detection of HCC at early stage with 100% sensitivity, however, these findings remain to be validated on bigger cohort (Cohen *et al.*, 2018).

Promising results have been also obtained for methylation-based approaches. Diagnostic panel based on 6 markers detected HCC at sensitivity 75-100% depending on stage and

outperformed AFP with AUC = 0.94 compared to 0.80 for AFP (Kisiel *et al.*, 2019).

Another study demonstrated that 5hmC profiles in gene panel can detect HCC with better performance compared to AFP (Cai *et al.*, 2019).

To conclude, previous studies suggest that early detection of HCC based on cfDNA methylation is promising approach, however, more research is needed in the field to develop more specific and sensitive assay with clear translational potential.

1.3 Overview of epigenetic tools for DNA methylation analysis

Discovery of new methylation biomarkers requires technologies that enable sensitive and specific mapping of 5mC and 5hmC. Various methods are currently available for detection and mapping of cytosine modifications (Figure 1-10), however, none of them are free from limitations and further technological improvements are necessary for more accurate and comprehensive analysis of 5mC and 5hmC in the genome, especially in challenging, low input clinical samples including cfDNA. In this chapter I reviewed key technologies available for DNA methylation analysis and discussed their strengths and weaknesses.

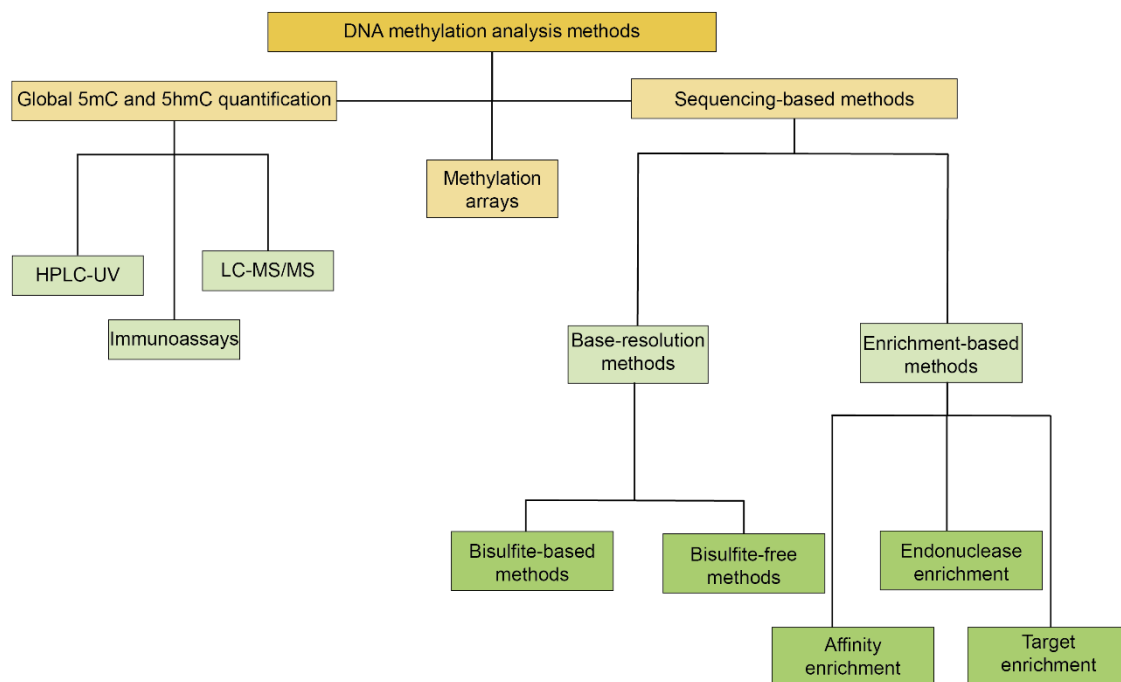


Figure 1-10. Classification of the most common approaches in DNA methylation analysis.

1.3.1 Global analysis of DNA methylation

Analysis of modified bases on a global scale is a common way to detect and quantify DNA modifications and was particularly popular before the development of next generation sequencing technologies.

Chromatography-based methods including High Performance Liquid Chromatography-Ultraviolet (HPLC-UV) (Armstrong *et al.*, 2011; Roberts, Raner and Isaacs, 2018) and Liquid chromatography coupled with tandem mass spectrometry (LC-MS/MS) (Song *et al.*, 2005; Ito *et al.*, 2011; M. Guo *et al.*, 2017) have been widely applied to analyse cytosine modifications in various DNA samples. To note, thin layer chromatography (TLC) followed by HPLC and MS enabled identification of 5hmC in 2009 (Kriaucionis and Heintz, 2009) and initial studies on TET-dependent DNA demethylation pathway (Ito *et al.*, 2011).

The above-mentioned techniques involve hydrolysis of DNA, followed by separation and detection of single nucleosides by UV or mass spectroscopy. These methods are able to specifically detect not only 5mC and 5hmC but also lowly abundant cytosine variants 5fC and 5caC (Ito *et al.*, 2011; M. Guo *et al.*, 2017; Roberts, Raner and Isaacs, 2018). LC-MS/MS is considered as a gold-standard to study epigenetic modifications at the global level owing to its high sensitivity. When coupled with electrospray ionisation LC-MS/MS was able to detect 5mC from as little as 4 ng of DNA with a limit of detection as low as 0.2 fmol (Song *et al.*, 2005).

Global level of cytosine modification can be also measured by immuno-based assay such as enzyme-linked immunosorbent assay (ELISA) or its modified version enhanced enzyme-based immunoassay (EEIA). ELISA was already applied for detection of 5mC (Kremer *et al.*, 2012) and 5hmC (Li and Liu, 2011) in human tissues. ELISA-based detection and quantification of 5mC could be performed from 10 ng of DNA with high accuracy. In turn, 5hmC assay could detect 0.2-10 ng that from 200 ng DNA sample. Assay for simultaneous analysis of 5mC, 5hmC and 5caC have been also developed and demonstrated limit of detection (LOD) values 3.802 pg for 5mC, 0.543 pg for 5hmC, 0.209 pg for 5caC that enabled to detect 5mC and 5hmC from as low as 100 ng while 5caC from 200 ng (Chowdhury *et al.*, 2014).

Global analysis of DNA methylation is convenient way to analyse new DNA modifications and capture global changes in modifications profiles. MS-based tools are highly sensitive and can be applied even to low input clinical sample but at the same time their wide use is limited by the requirement of special equipment and relatively high costs (M. Guo *et al.*, 2017). Immunoassays are more user-friendly and easier to be performed in non-specialised research settings, but they exhibit lower sensitivity and provide less accurate quantification compared to the LC-MS/MS. Both types of assays are able only

to capture global changes in methylation but do not provide any sequence information. Therefore, they may miss subtle yet biologically significant local changes in 5mC and 5hmC profiles.

1.3.2 Enrichment-based methods

Although whole-genome methylome sequencing methods give the most complete picture of epigenetic modifications in DNA, they are also very expensive (Plongthongkum, Diep and Zhang, 2014). A cheaper alternative are enrichment-based methods for methylation mapping that involve enrichment of modified DNA fragments by restriction digestion (Meissner *et al.*, 2005; Wang *et al.*, 2013) or affinity enrichment (Song *et al.*, 2011; Taiwo *et al.*, 2012; Zhang *et al.*, 2013) before sequencing. As only around 5-6% of all cytosines in the genome are modified, even up to 80% of sequencing reads generated by whole-genome methylation sequencing may not carry any methylation information. Excluding these redundant regions from analysis by an enrichment step helps to restrict sequencing only to the most informative parts of the genome and significantly reduce costs (Ziller *et al.*, 2013; Plongthongkum, Diep and Zhang, 2014; Breiling and Lyko, 2015).

One approach to enrich DNA fragments for methylome analysis are endonuclease-enrichment based assays such as Methylation-Sensitive Restriction Enzyme Sequencing (MRE-seq) (Maunakea *et al.*, 2010). In MRE-seq DNA is digested by a cocktail of methylation-sensitive restriction enzymes before size-selection and construction into sequencing library (Figure 1-11). As only unmethylated CpG sites can be digested, they are enriched at the ends of cleaved fragments and thus can be used to reveal the location of unmodified sites and estimating relative methylation (Plongthongkum, Diep and Zhang, 2014).

MRE-seq

DNA digestion with methylation-sensitive restriction enzyme

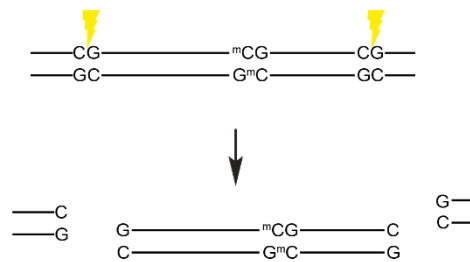


Figure 1-11. Simplified overview of MRE-seq protocol.

Although MRE-seq enables the significant reduction of methylation sequencing costs, it can only detect CpG sites in specific sequence contexts and in the regions with relatively high density of CpG sites. Given that this method covers only around 6% of all CpG sites in the human genome, it generates incomplete methylation maps that leave the majority of CpG sites in the genome uncharacterized (Maunakea *et al.*, 2010).

Affinity enrichment methods can be a more efficient way to capture methylated DNA before sequencing. These methods involve enrichment of 5mC or 5hmC by specific antibodies (Maunakea *et al.*, 2010; Ficz *et al.*, 2011; Taiwo *et al.*, 2012), proteins (Aberg *et al.*, 2012) or chemical labelling (Song *et al.*, 2011). In this way only DNA fragments containing modifications of interest are constructed into a library and sequenced, which significantly reduces sequencing costs.

Methylated DNA immunoprecipitation sequencing (MeDIP-seq) involves enrichment of methylated fragments by specific antibody (Figure 1-12) and can cover around 86% of all CpG sites in the human genome (Maunakea *et al.*, 2010; Robinson *et al.*, 2010; Taiwo *et al.*, 2012).

MeDIP

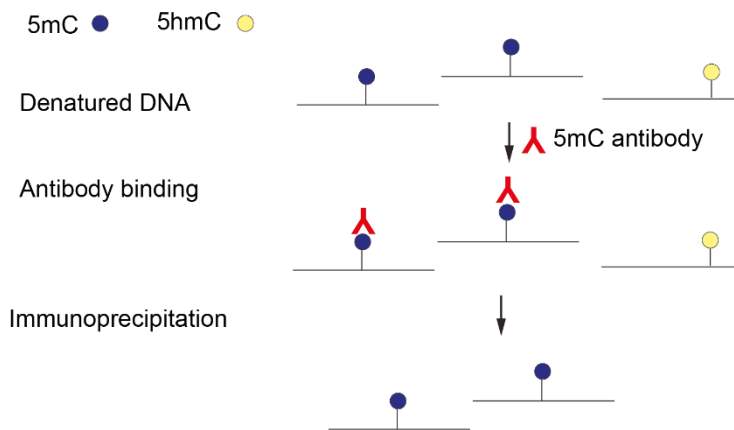


Figure 1-12. Simplified overview of MeDIP protocol.

Initial MeDIP-seq protocols required relatively high amount of DNA, however, recently immunoprecipitation protocols have been optimised to work with minute amount of DNA and was applied for methylation analysis in cfDNA clinical samples (Shen *et al.*, 2018). Similar methods for 5hmC enrichment have been also developed extending application of antibody-based DNA immunoprecipitation to other cytosine variants (Thomson *et al.*, 2013).

Alternatively, cytosine modifications can be captured by binding proteins e.g. by methyl-CpG binding domain (MBD) protein or by Thymine dioxygenase JBP1 (Plongthongkum, Diep and Zhang, 2014). In MBD-seq 5mC is captured by MBD protein that recognizes and binds methylated cytosines specifically in a CpG context (Robinson *et al.*, 2010; Nair *et al.*, 2011). A similar enrichment method was developed to capture glucose-modified 5hmC by JBP1 proteins (Thomson *et al.*, 2013).

Notably, antibody and protein-based enrichment methods significantly reduce sequencing costs while still covering over 80% of CpG sites in the genome (Maunakea *et al.*, 2010; Plongthongkum, Diep and Zhang, 2014). However, they are strongly affected by affinity and specificity of antibodies or proteins used for enrichment. Numerous studies have already demonstrated that enrichment can be biased by CpG-density and

copy number variations (CNVs). Additionally, unspecific binding can occur in short tandem repeats (STRs) overestimating the modification level in those regions (Robinson *et al.*, 2010; Nair *et al.*, 2011; Thomson *et al.*, 2013; Lentini *et al.*, 2018).

Selective chemical labelling and capture of modified DNA can be a more specific way of enrichment (Thomson *et al.*, 2013). For instance, hMeSeal enables enrichment of hydroxymethylated DNA by enzymatic labelling of 5hmC sites with azide-modified glucose followed by biotin transfer in click-chemistry reaction. Biotinylated fragment can be subsequently captured on streptavidin beads and sequenced (Figure 1-13) (Song *et al.*, 2011).

hMeSeal

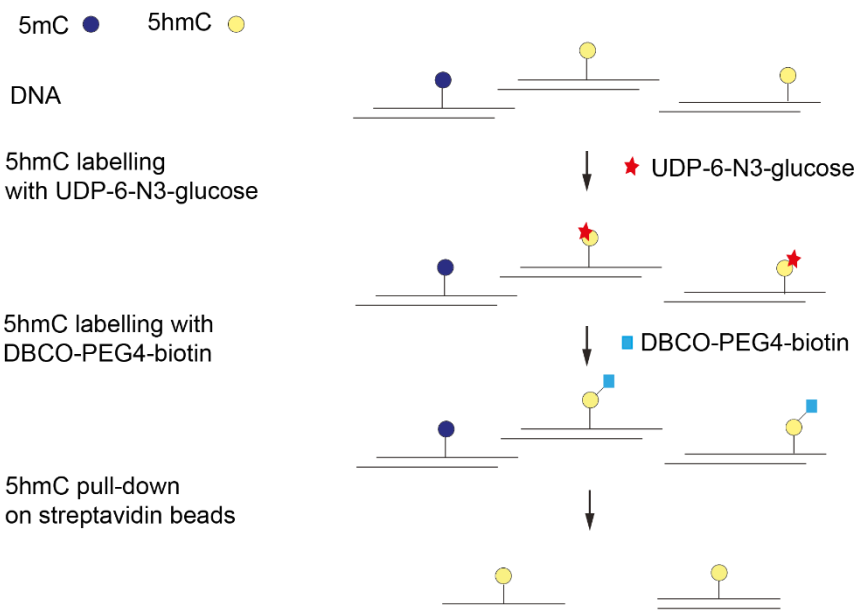


Figure 1-13. Simplified overview of hMeSeal protocol.

A similar method was developed for labelling 5mC by combining TET oxidation with transfer of azide-glucose (Zhang *et al.*, 2013). 5hmC mapping by hMeSeal-seq strongly correlates with results produced by hMeDIP-seq, however, it is less affected by unspecific binding to STRs (Thomson *et al.*, 2013). hMeSeal method works with as low as 10 ng of DNA and was already successfully applied to study cancer-specific hydroxymethylation signatures in cfDNA (Song *et al.*, 2017).

Different approach to enrichment sequencing was proposed by Kriukiene *et al.* who developed Methyltransferase-directed Transfer of Activated Groups Sequencing (mTAG-seq) based on labelling of unmodified CpG sites instead of capturing modified cytosines (Kriukiene *et al.*, 2013). In their method unmodified CpG sites are aminoalkylated by genetically engineered M. SssI methyltransferase and Ado-6-amine, labelled with biotin, captured on streptavidin beads and sequenced (Figure 1-14). This approach enables better characterization of genomic regions with low and intermediate CpG density when compared to MeDIP-seq or MBD-seq methods that are biased towards regions with high density of CG dinucleotides.

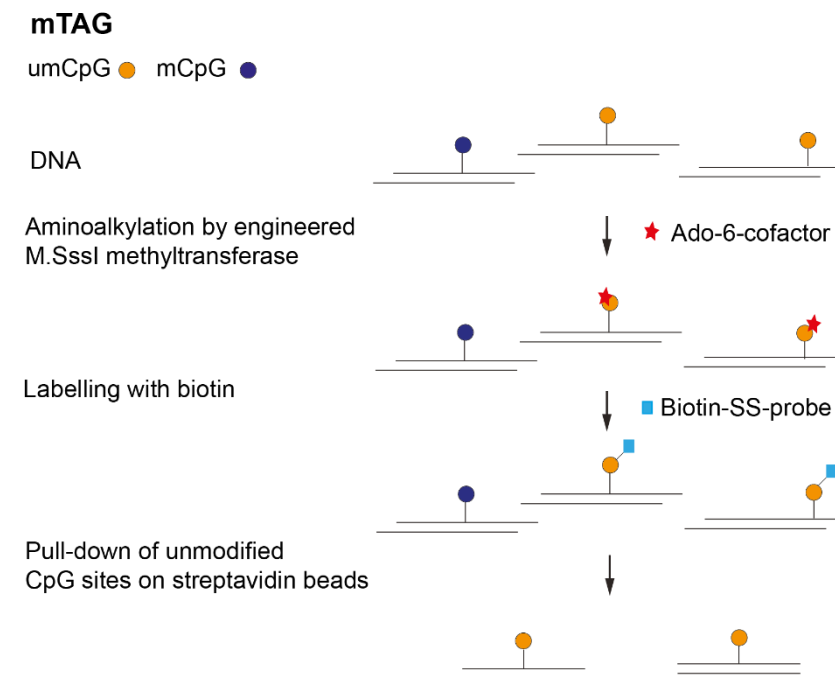


Figure 1-14. Overview of mTAG labelling. umCpG – unmodified CpG site. mCpG – modified CpG site.

Although, affinity enrichment methods demonstrated their invaluable usefulness in elucidating role of cytosine modification, their resolution is limited to 150-1000 bp and they lack base-resolution information. Additionally, they are not fully quantitative and estimate only relative modification levels (Maunakea *et al.*, 2010; Thomson *et al.*, 2013; Plongthongkum, Diep and Zhang, 2014). Reduced Representation Bisulfite Sequencing

(RRBS) was developed to address this problem and combine cost-effectiveness of enrichment approaches with single base-resolution. This method involves restriction digestion-based enrichment of CpG-rich regions followed by bisulfite treatment (Figure 1-15) (Meissner *et al.*, 2005).

RRBS



Figure 1-15. Simplified scheme of RRBS protocol.

Although RRBS still provides modification levels for individual cytosines, it covers only a small fraction of cytosines mostly within CpG-dense region and is restricted to specific sequence contexts strictly depending on restriction enzymes used for DNA digestion (Plongthongkum, Diep and Zhang, 2014). However, due to relatively low costs RRBS has been widely used in the epigenetic studies and the protocol was further improved to work with low DNA input down to single cell level (Smallwood *et al.*, 2014).

Target-enrichment methods can be considered a separate type of enrichment approach. They are focused on methylation analysis in single locus or set of loci (Plongthongkum, Diep and Zhang, 2014). Target enrichment can be achieved by PCR amplification or Hybridization Capture-based Enrichment of bisulfite-converted DNA. This allows to test methylation status only in loci of interest with low costs, however can be sensitive to PCR-bias and PCR-errors (Guibert *et al.*, 2018; M. C. Liu *et al.*, 2020).

Taken together, the main advantage of enrichment-based methods for methylome analysis are reduced costs. However, it is crucial to remember, that they provide less

comprehensive picture of DNA modifications compared to whole-genome sequencing approaches.

1.3.3 Base-resolution methods for whole-genome methylation analysis

Development of next generation sequencing methods opened new opportunities to study the epigenome. However, during standard whole-genome sequencing approaches cytosine modifications are not distinguishable from unmodified cytosine. Therefore, additional steps are needed to enable selective detection of DNA modifications.

The most common way and current gold standard to sequence DNA methylation at single-base resolution is bisulfite sequencing (BS-seq) (Frommer *et al.*, 1992; Susan *et al.*, 1994). Bisulfite treatment relies on ability of bisulfite treatment to selectively deaminate unmodified cytosines to uracil which is subsequently read as thymine during PCR amplification while retaining 5mC and 5hmC intact (Figure 1-16, a-b). Recent studies suggest that 5fC and 5caC behave as unmodified cytosines and became converted by bisulfite, thus standard bisulfite sequencing gives selective read-out of 5mC and 5hmC at the single-base resolution (He *et al.*, 2011; Wu, Wu and Zhang, 2016).

Standard BS-seq has been modified to distinguish between 5mC and 5hmC. In Tet-assisted bisulfite sequencing (TAB-seq) 5hmC is selectively blocked by UDP-glucose in reaction catalysed by T4-beta-glucosyltransferase and is resistant to both TET oxidation and bisulfite reaction while 5mC is oxidised by TET enzyme to 5fC/5caC and can be converted to uracil upon bisulfite treatment. After PCR amplification 5hmC is read as cytosine while other cytosine variants are read as thymine (Figure 1-16, b). TAB-seq provides specific, base-resolution read-out of 5hmC in the genome and 5mC positions can be indirectly detected by comparing of TAB-seq results with standard bisulfite sequencing (Yu *et al.*, 2012).

Another method based on bisulfite-treatment is as Oxidative Bisulfite Sequencing (oxBS-seq) where 5hmC is chemically oxidised by oxidising agent to produce 5fC prior bisulfite treatment. In this method only 5mC resists conversion and is read as C while other cytosine variants are converted and detected as T (Figure 1-16, b) (Booth *et al.*, 2012; Peng, Xia and Yi, 2016).

Although BS-seq and related methods are considered as a gold standard in methylome analysis, they have serious disadvantages. First, bisulfite conversion involves harsh chemical treatment that denatures DNA, leads to DNA fragmentation and can degrade up to 95% of original input, which results in uneven coverage and limits its application to low input samples (Tanaka and Okamoto, 2007). Second, it generates libraries with low sequence complexity compared to standard whole genome sequencing (Olova *et al.*, 2018). In standard, unconverted genome all nucleotides are present at similar proportion around 25%. Given that most of cytosines in the genome are unmethylated and converted to T in bisulfite treatment, only around 5% of all cytosines resist conversion and thus cytosines constitute of around 1% of all nucleotides in converted genome. Overrepresentation of thymines and adenines and underrepresentation of cytosines and guanines in final bisulfite libraries results in a significant drop in the sequence complexity and creates technical problems during Illumina sequencing and data pre-processing. Due to mentioned disadvantages, new technologies independent on bisulfite treatment are required to facilitate future genome-wide epigenetic studies, especially in low input samples.

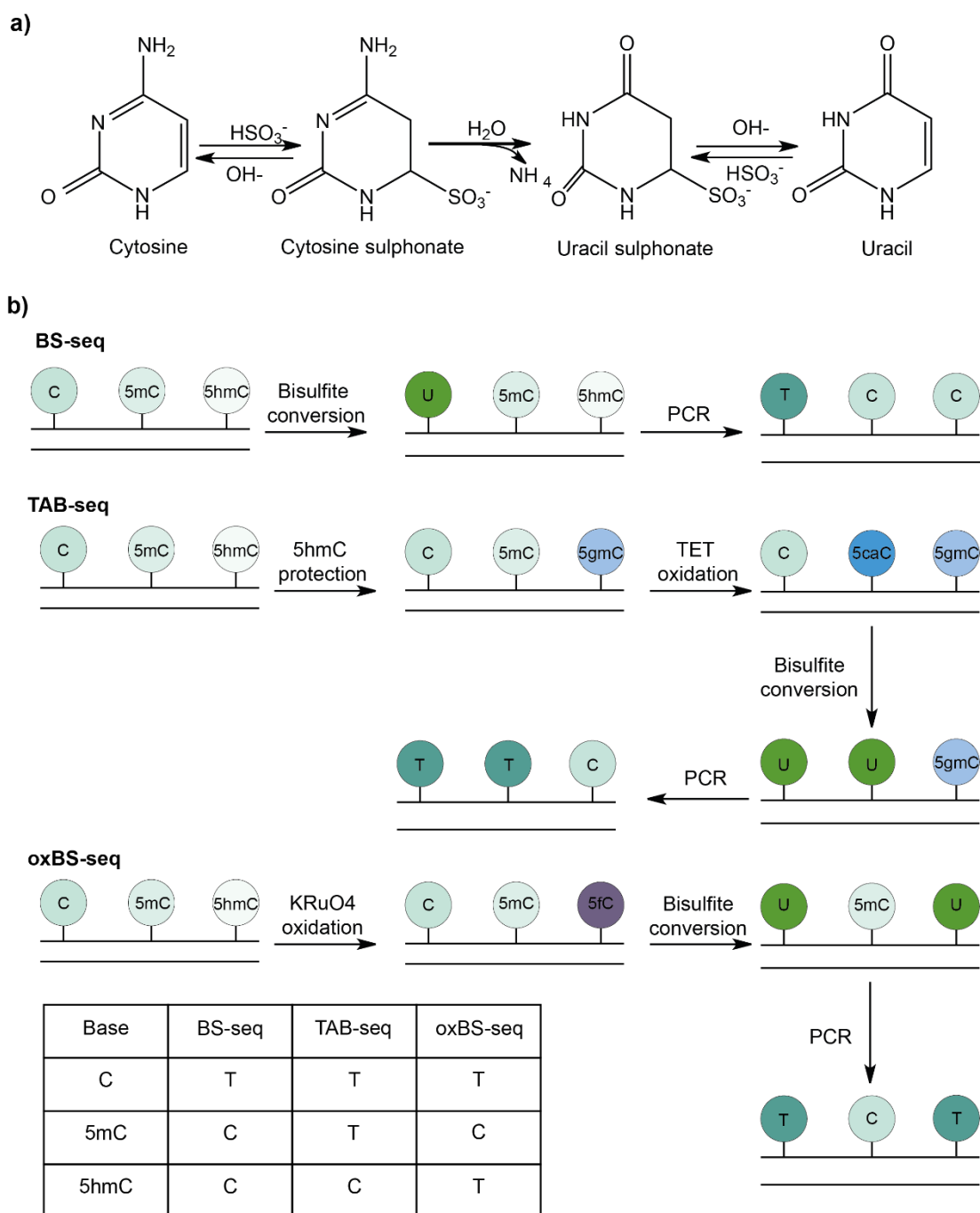


Figure 1-16. Summary of bisulfite-based methods for 5mC and 5hmC sequencing. (a) Mechanisms of bisulfite conversion of unmodified cytosine. Bisulfite reaction involves three steps. In first step, sulphonylation, sodium bisulfite is added to cytosine to generate cytosine sulphonate. In second step, cytosine sulphate is deaminated to uracil sulphonate. In the last step uracil sulphonate is desulphonated to uracil, the final product of conversion. (b) Comparison of bisulfite-based methods for cytosine modification sequencing. Final readout of the methods for C, 5mC and 5hmC is shown in the table. Abbreviations: C - cytosine; 5caC - 5-carboxycytosine; 5fC - 5-formylcytosine; 5gC - β -glucosyl-5-hydroxymethylcytosine; 5hmC - 5-hydroxymethylcytosine; 5mC - 5-methylcytosine; T - thymine; U - uracil.

Recently, new bisulfite-free approaches based on enzymatic deamination of cytosines by Apolipoprotein B mRNA editing catalytic polypeptide-like (APOBEC) enzyme were developed (Schutsky *et al.*, 2018; Vaisvila *et al.*, 2019). APOBEC enzymes deaminate unmodified cytosines as well as 5mC and 5hmC, therefore deamination had to be coupled with additional treatment to be applied for methylome sequencing.

In APOBEC-coupled epigenetic sequencing (ACE-seq) glucose moiety is enzymatically transferred to hydroxymethylated sites before APOBEC treatment therefore 5hmC remains protected from deamination. In ACE-seq 5hmC is read as C in the final sequencing while unmodified cytosines and 5mC are read as thymines (Figure 1-17, a) (Schutsky *et al.*, 2018). This approach has been further modified to detect both 5mC and 5hmC. This was achieved by TET-mediated oxidation of 5mC and 5hmC to 5caC so they remain resistant to deamination (Vaisvila *et al.*, 2019). Similar to bisulfite sequencing, this method detects 5mC and 5hmC as cytosine while unmodified cytosines as thymines (Figure 1-17, b). Although APOBEC-based methods are milder than bisulfite-based approaches and do not cause extensive DNA degradation, they are still based on conversion of unmodified cytosines, therefore, similar to bisulfite sequencing, produce low complexity libraries. Moreover, as both 5fC and 5caC are resistant to enzymatic deamination they are detected as 5mC and 5hmC, however, due to their low level in the genome it can be potentially neglected.

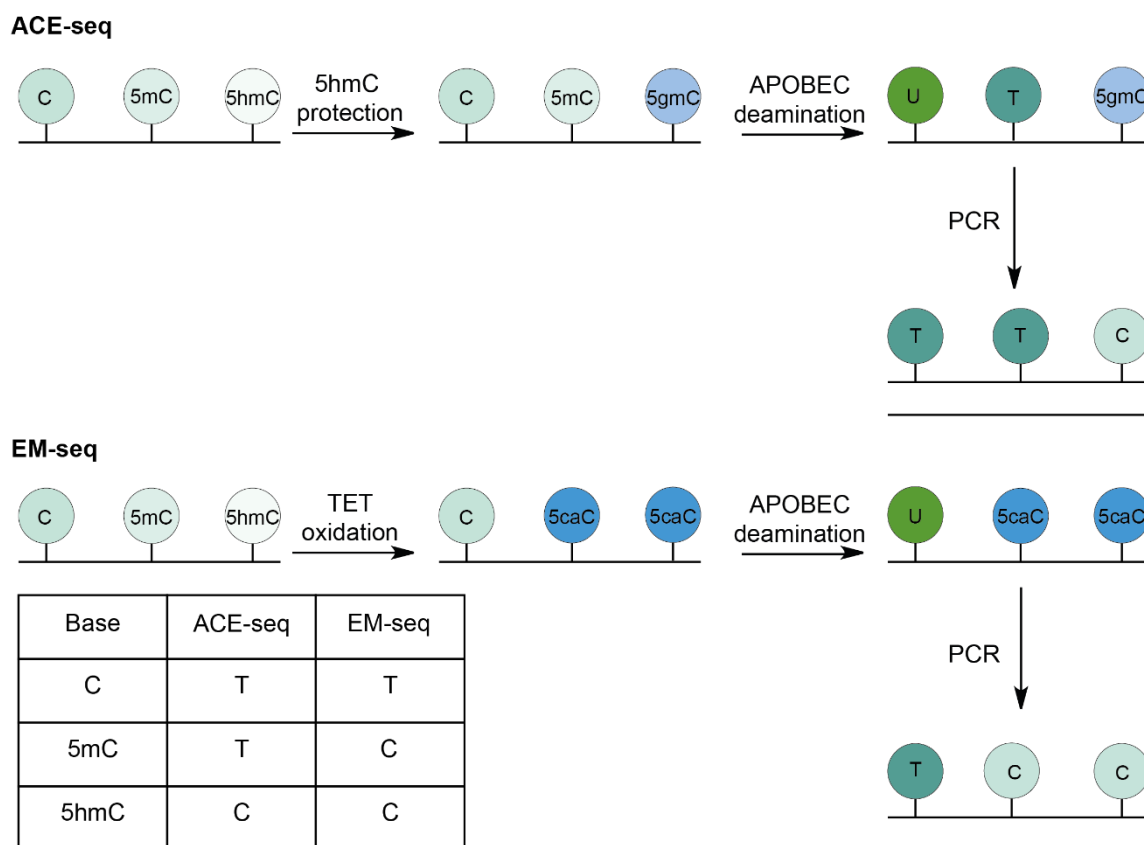


Figure 1-17. Summary of APOBEC-based DNA methylation sequencing methods. Final readout of the methods for C, 5mC and 5hmC is shown in the table. Abbreviations: C - cytosine; 5caC - 5-carboxylcytosine; 5gmC - β -glucosyl-5-hydroxymethylcytosine; 5hmC - 5-hydroxymethylcytosine; 5mC - 5-methylcytosine; T - thymine; U - uracil.

To conclude, whole-genome, base-resolution methylome and hydroxymethylome sequencing methods are the most comprehensive and sensitive way to study epigenetic modifications in DNA, however, their relatively high costs limit their application in some studies. Also, this category is strongly dominated by methods involving bisulfite treatment, which highlights the need for development of novel, bisulfite-free approaches to address all technical limitations of the current technologies.

1.3.4 Methylation arrays

Methylation arrays are another way for cost-effective analysis of DNA methylation (Plongthongkum, Diep and Zhang, 2014). They have been most widely used in epigenetic research as they are relatively cheap, high-throughput and easy-to-use.

One of the most common methylation arrays are Illumina Infinium 450K and related chips that estimate methylation levels in predefined CpG and non-CpG positions. This technology is based on hybridization of bisulfite-converted DNA to the complementary probe that ends with potential methylation site at 3' terminus. Methylation level at individual site is estimated based on the fluorescent signal generated during single-base extension of the probe with labelled nucleotide (Pidsley *et al.*, 2016).

In addition to being user-friendly and cheap, methylation arrays provide reproducible and reliable methylation estimates that remain in good agreement with BS-seq (Pidsley *et al.*, 2016). As they cover pre-selected sites mostly with previously confirmed biological significance, analysis is focus on the most crucial regions in DNA. On the other hand, as covered sites constitute only small proportion of the genome, arrays may miss important and biologically significant methylation changes that occur outside regions represented on the chip. Additionally, existing methylation arrays involve DNA treatment with bisulfite which itself introduces DNA damage, reduces sequence complexity and can contribute to unspecific hybridization of target DNA to the probes (Wilhelm-Benartzi *et al.*, 2013; Plongthongkum, Diep and Zhang, 2014; Pidsley *et al.*, 2016).

1.3.5 Third generation sequencing methods for methylome analysis

Recent development of third generation sequencing methods opened new possibilities for direct detection of methylated sites without the need for chemical conversion (van Dijk *et al.*, 2018). One of the most known problems existing in short read Illumina sequencing are problems with mapping of low complexity, repetitive regions. Additionally, whole-genome amplification, which is common step in most of methods for epigenetic sequencing, can introduce sequence-dependent bias, especially in difficult to amplify GC-rich regions (Robinson *et al.*, 2010; Olova *et al.*, 2018). Direct, long-read sequencing

can potentially help to address these problems and offer significant improvement in mapping DNA modifications.

Single-molecule, real-time sequencing (SMRT-seq) is one of the first 3rd generation sequencing technologies. In this method polymerase copies single DNA molecule and replication is monitored based on the signal emitted during incorporation of fluorescently labelled nucleotides. The intensity and duration of this signal are used for quantification of polymerase speed and identification of the added base. Incorporation of a new nucleotide opposite to epigenetic marks e.g. 5mC by polymerase requires more time and causes an increase in time interval between fluorescent signals. This is used for distinction and characterization of modified sites (van Dijk *et al.*, 2018; Wenger *et al.*, 2019).

Another long-read sequencing method is Oxford Nanopore technology that utilizes nanopores for identification of nucleotides in DNA. Nanopores are protein channels incorporated into the membrane. Each nanopore is in a separate compartment filled with ionic solution, which causes ionic current to move through the pore. DNA molecules moving through the pore generates changes to the ionic current and these changes can be then translated into the sequence information (Jain *et al.*, 2018; van Dijk *et al.*, 2018).

Both SMRT-seq and Oxford Nanopore sequencing technologies enable PCR-free, direct detection of modified bases with long read-sequencing, which can improve epigenetic analysis especially in the low-mappability regions. They can also help to avoid sequence-dependent or amplification dependent biases caused by whole-genome amplification or chemical conversion used in the traditional epigenetic sequencing methods.

On the other hand, these technologies are still under development. Nanopore sequencing still exhibits high error rate ~ 10% that may be too high for detection of reliable modified bases that exist at low abundance in the genome (Rand *et al.*, 2017). SMRT-seq already

improved accuracy in base calling to the level comparable with Illumina sequencing, however, it requires high coverage and high input amount to produce reliable methylation read-out (Wenger *et al.*, 2019). Due to those limitations, long-read sequencing technologies require further improvement before replacing short-read sequencing-based methylation analysis.

1.3.6 Concluding remarks

As demonstrated above, multiple methods and approaches exist to map DNA methylation and hydroxymethylation. All of them bring their own strengths and weaknesses, which must be carefully considered during experimental design.

Briefly, global methylation analysis by chromatography-based methods or Mass Spectrometry played a crucial role in the discovery of epigenetic modifications, but they only capture global changes in methylation. Affinity-based methylation sequencing methods allow for cost-effective methylation analysis and are less damaging than bisulfite-based methods, but do not provide base resolution methylation information. Base-resolution methylation sequencing methods provide the most comprehensive mapping of 5mC genome-wide but most of them require bisulfite conversion, which introduces extensive damage to input DNA and generates low complexity sequencing libraries. Methylation arrays are widely used in epigenetic studies owing to low costs and robustness but cover only a small proportion of the genome. Recently developed long-read methylation sequencing methods offer a novel approach for methylation sequencing but are still associated with high error rate.

As none of the existing technologies are demonstrably free from serious drawbacks, further developments in the field are still required to provide tools for more accurate and

less biased analysis of DNA modifications, to facilitate methylome analysis in low input samples.

2 Research objectives

Increasing understanding of cancer epigenetics and involvement of DNA methylation in cancer development and progression opened new directions to use DNA methylation marks as cancer biomarkers, especially in combination with cfDNA analysis (van der Pol and Mouliere, 2019; Brancaccio *et al.*, 2020).

Current knowledge suggest that methylation analysis can outperform genetic markers in early detection liquid biopsy assays, however, most of current methylation biomarkers have not been transferred to the clinical practice, suggesting that more research focused on cancer methylation, especially in cfDNA are needed (Neal *et al.*, 2015; Koch *et al.*, 2018). Moreover, simultaneous analysis of epigenetic and genetic changes in a single sequencing assay could facilitate development of more sensitive, multianalyte tests for cancer detection. Unfortunately, base-resolution methylation sequencing is currently almost exclusively limited to bisulfite-related methods that are not easily applicable to cfDNA analysis and do not provide reliable information about other cfDNA features such as fragmentation profiles. Therefore, new bisulfite-free sequencing methods overcoming the limitations of BS-seq and enabling comprehensive analysis of epigenetic and genetic changes in cfDNA analysis are required.

The main goal of my DPhil project was development of bisulfite-free methods for methylome sequencing and their application for methylation analysis in the clinical cfDNA samples.

The specific goals of my project were:

- Development of an enrichment-based 5mC mapping method (cfMeSeal) for cfDNA sequencing. Enrichment methods are well-accepted and cost-effective way for mapping 5mC and 5hmC in the genome, however, most of 5mC

enrichment approaches are based on the antibody enrichment and may be affected by specificity issues. The main aim of this part was to develop cfMeSeal method using highly specific, chemical labelling instead of immunoprecipitation for 5mC enrichment and to validate its performance in sequencing of cfDNA from cancer and healthy subjects (Chapter 3).

- Development of a bisulfite-free, base resolution method, Tet-assisted pyridine borane sequencing (TAPS), for methylation sequencing to overcome limitations of the current gold standard, BS-seq (Chapter 4).
- Development of cfDNA TAPS protocol and its application for genome-wide methylation profiling in cfDNA samples in healthy, PDAC and HCC. The aim of this project was to test potential of TAPS to provide genome-wide, multidimensional information about cfDNA characteristics including methylation and fragmentation signatures and to assess their predictive potential (Chapter 5).
- Development and validation of endonuclease enrichment TAPS (eeTAPS) protocol for methylome sequencing. Although TAPS methods showed advantages compared to BS-seq, it is a whole-genome sequencing approach and thus is relatively expensive. The aim of this part was to develop enrichment-based modified TAPS protocol for cheaper, cost-effective methylome sequencing (Chapter 6).

List of publications associated with this dissertation are listed in Appendix 2. Most of Chapter 4 was published in Liu *et al.* (2019). Manuscripts based on Chapter 5 and Chapter 6 are under preparation.

Have no fear of perfection; you'll never reach it.

Maria Skłodowska-Curie

3 Cell-free DNA MeSeal method for 5mC enrichment

3.1 Introduction

Since DNA methylation is frequently affected in malignancies, methylation-based biomarkers are considered as promising targets for the development of novel tools for cancer detection and monitoring. Moreover, detection of cancer-specific methylation marks in cell-free DNA (cfDNA) would offer a new, minimally invasive way to diagnose cancer (Berdasco and Esteller, 2018; Koch *et al.*, 2018; Brancaccio *et al.*, 2020).

Unfortunately, epigenetic analysis of cfDNA is technologically challenging. CfDNA level in plasma is low and rarely exceeds 10 ng per 1 mL of plasma (Wan *et al.*, 2017). Due to low amount of the starting material and demand for higher sensitivity to detect a low tumour fraction, most methylation studies in the cfDNA field are limited to pre-defined regions and loci. On the other hand, genome-wide methylation analysis with bisulfite sequencing offers the most comprehensive way to map 5-methylcytosine (5mC) in cfDNA. However, this approach is expensive and quality of data is affected by DNA damage introduced by the chemical treatment (Plongthongkum, Diep and Zhang, 2014; Worm Ørntoft *et al.*, 2017).

Affinity-based enrichment methods are cost-effective alternatives to the whole genome methylome sequencing while they are still able to cover a substantial fraction of CpG sites in the genome. The most common enrichment method for methylation sequencing is Methylated DNA immunoprecipitation coupled with next generation sequencing (MeDIP-seq) where modified DNA fragments are captured by a 5mC-specific antibody (Maunakea *et al.*, 2010; Robinson *et al.*, 2010; Taiwo *et al.*, 2012). The original MeDIP-

seq protocol required high DNA input and thus was not suitable for cfDNA analysis but recently MeDIP-seq method optimised to work with cfDNA samples has been published (Shen *et al.*, 2018; Nassiri *et al.*, 2020; Nuzzo *et al.*, 2020).

On the other hand, antibodies can recognize and bind to short tandem repeats (STRs) causing a false positive methylation signal from those regions (Thomson *et al.*, 2013; Lentini *et al.*, 2018). Alternative and potentially more specific approaches to capture modified DNA are methods based on chemical and/or enzymatic labelling and pull-down of modified bases. For example, 5hmC can be chemically labelled, biotinylated and captured on streptavidin beads in hMeSeal method and some studies already suggested that this strategy provides more specific enrichment compared to immunoprecipitation (Song *et al.*, 2011; Thomson *et al.*, 2013; Lentini *et al.*, 2018). Moreover, hMeSeal method is a mild and non-damaging reaction and has been already successfully applied for hydroxymethylation sequencing in cfDNA demonstrating good potential to identify cancer-specific 5hmC patterns (Song *et al.*, 2017). Development of an analogous method for specific mapping of 5mC would enable comprehensive, enrichment-based analysis of methylation in cfDNA and would offer an alternative to the currently existing methods in the field.

3.2 Objectives

Enrichment methods for 5mC mapping in cfDNA based on chemical labelling could offer a convenient approach to analyse methylation signatures and facilitate epigenetic studies in circulating DNA. Previously developed hMeSeal method has been proved to capture aberrant hydroxymethylation signatures in cfDNA from cancer patients (Song *et al.*, 2017). We sought to develop a similar protocol for 5mC mapping in cfDNA. 5mC labelling method that combines TET oxidation and labelling with azide-modified glucose

was already published, however, the protocol has been not expanded to low input samples (Zhang *et al.*, 2013). The main goal of this project was to optimise the published 5mC labelling protocol to cfDNA analysis and to test its feasibility to detect cancer-associated aberrant cfDNA methylation signatures in a pilot project.

3.3 General design of cfMeSeal method

cfMeSeal method for 5mC enrichment in cfDNA samples was designed based on the method previously described in literature for high input genomic DNA samples (Zhang *et al.*, 2013). In the first step, 5hmC is blocked by UDP-glucose in an enzymatic reaction catalysed by T4 β -glucosyltransferase (T4-BGT) (Figure 3-1, a). After blocking, hydroxymethylated sites cannot be labelled and enriched in the following steps, which enables specific detection of DNA methylation. In the next stage, 5mC is oxidised to 5hmC and labelled with azide-modified glucose in a one-pot reaction with NgTET1 and T4-BGT (Figure 3-1, b). Subsequently, biotin is added to azide-labelled DNA fragments in a click chemistry reaction. In the final step Illumina adapters are ligated, biotinylated DNA fragments are captured on streptavidin beads and the final library is amplified and sequenced. Complete cfMeSeal procedure is summarised in Figure 3-2.

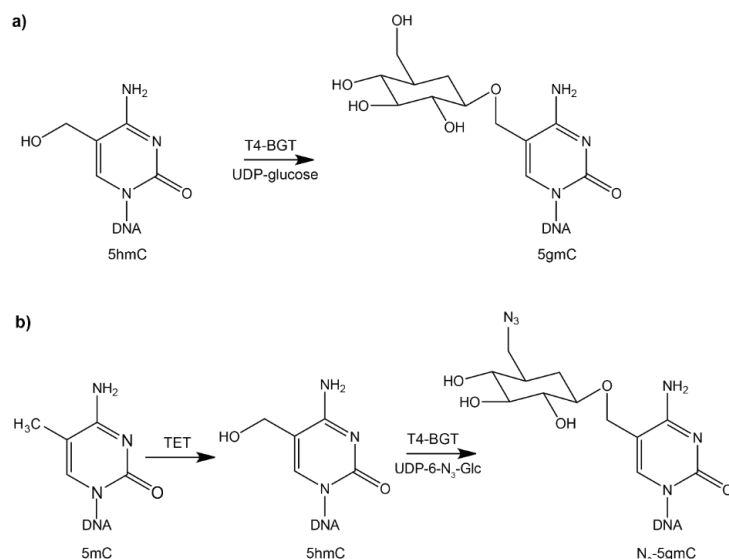


Figure 3-1. Chemical reactions in cfMeSeal method. (a) Selective blocking of 5hmC with UDP-glucose. (b) One-pot oxidation and labelling of 5mC. 5mC oxidation to 5hmC is followed by a transfer of UDP- N_3 -glucose. N_3 -5gmC can be then labelled with biotin in a click chemistry reaction and captured on streptavidin beads. 5gmC - glucosyl-5-hydroxymethyl-cytosine; UDP-6- N_3 .Glc - UDP-6-azide-glucose; N_3 -5gmC - 6- N_3 - β -glucosyl-5-hydroxymethyl-cytosine.

To evaluate specificity and efficiency of 5mC pull-down, set of 8 spike-in DNA control oligos was designed. This control set included unmethylated and hydroxymethylated oligos as well as DNA fragments containing different number of methylated CpG sites. Enrichment specificity could then be evaluated using PCR-based methods during method optimisation or based on the number of reads mapped to the control sequences in sequencing libraries.

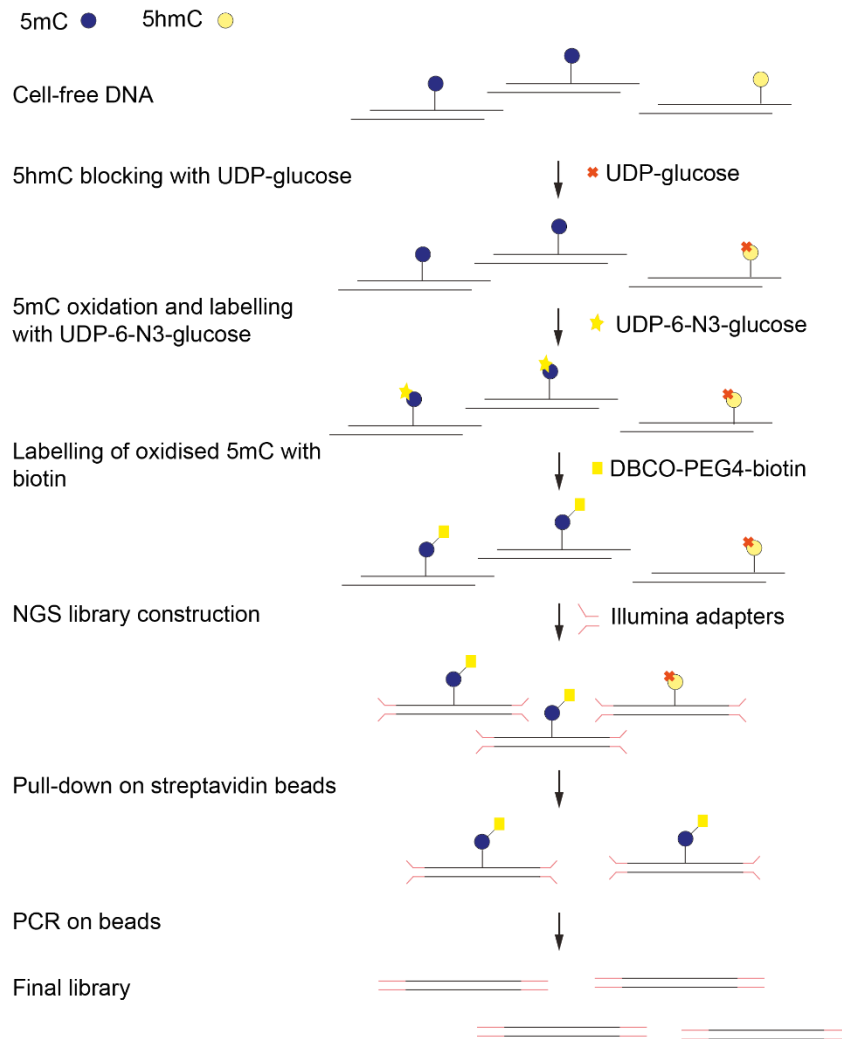


Figure 3-2. Overview of cfMeSeal method.

3.4 Non-mammalian, recombinant TET enzymes can be successfully expressed and purified from *E. coli*

Development of cfMeSeal method required an active TET enzyme for 5mC oxidation. We selected two TET enzymes that can be easily expressed in *E. coli*, *Naegleria guberi* TET1 (NgTET1) (Hashimoto *et al.*, 2014; Pais *et al.*, 2015) and *Coprinopsis Cinerea* TET (ccTET) (Zhang *et al.*, 2014) as alternatives to more laborious production of mammalian TET enzymes in insect cells (Yu *et al.*, 2012).

NgTET1 protein was expressed as an N-terminal His-tag protein and purified in 3-step protocol including Nickel-affinity chromatography, size-exclusion chromatography and ion exchange chromatography (Figure 3-3). In turn, ccTET was purified in 2-step protocol without the ion exchange chromatography step due to the lower protein yield (Figure 3-4). Both purifications yielded pure, non-degraded proteins with the major bands corresponding to the expected molecular weights.

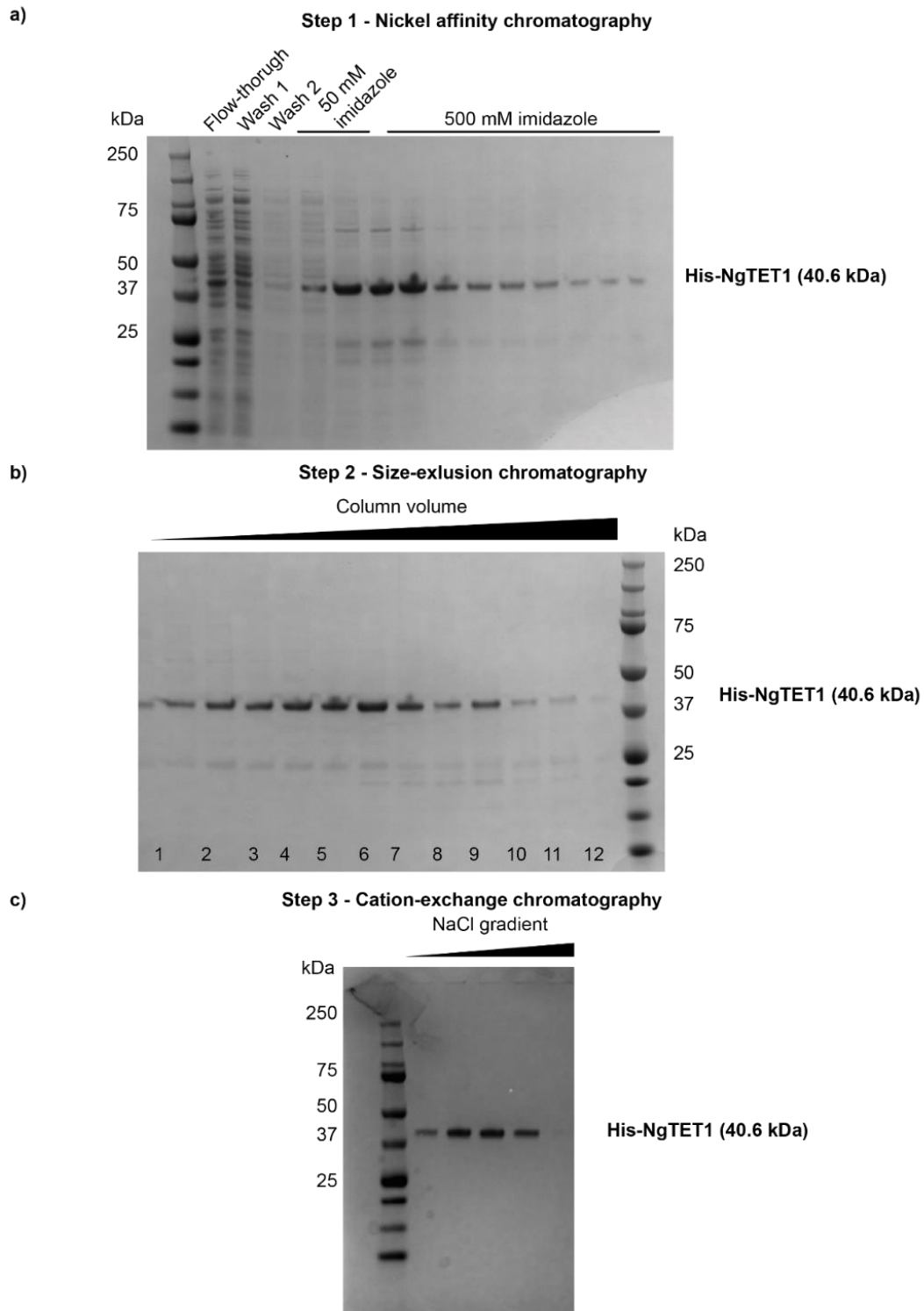


Figure 3-3. Representative, Coomassie-stained SDS PAGE electrophoresis gels for purification of His-NgTET1. (a) SDS-PAGE gel showing protein fractions after His-NgTET1 purification on Ni-NTA resins. Flow-through contained unbound bacteria proteins. Wash 1 and wash 2 with low imidazole concentration removed majority of contaminant proteins. His-tagged NgTET1 protein was then eluted after washing with the buffers containing 50 mM and 500 mM imidazole. All eluted fractions were used for size-exclusion chromatography. (b) SDS PAGE after second step of His-NgTET1 purification by size-exclusion chromatography. Fractions from lanes 1-9 were collected for final clean-up by cation-exchange chromatography. (c) SDS PAGE gel of pure His-

NgTET1 protein fractions after cation-exchange chromatography. All protein-containing fractions were collected, concentrated and used in further assays.

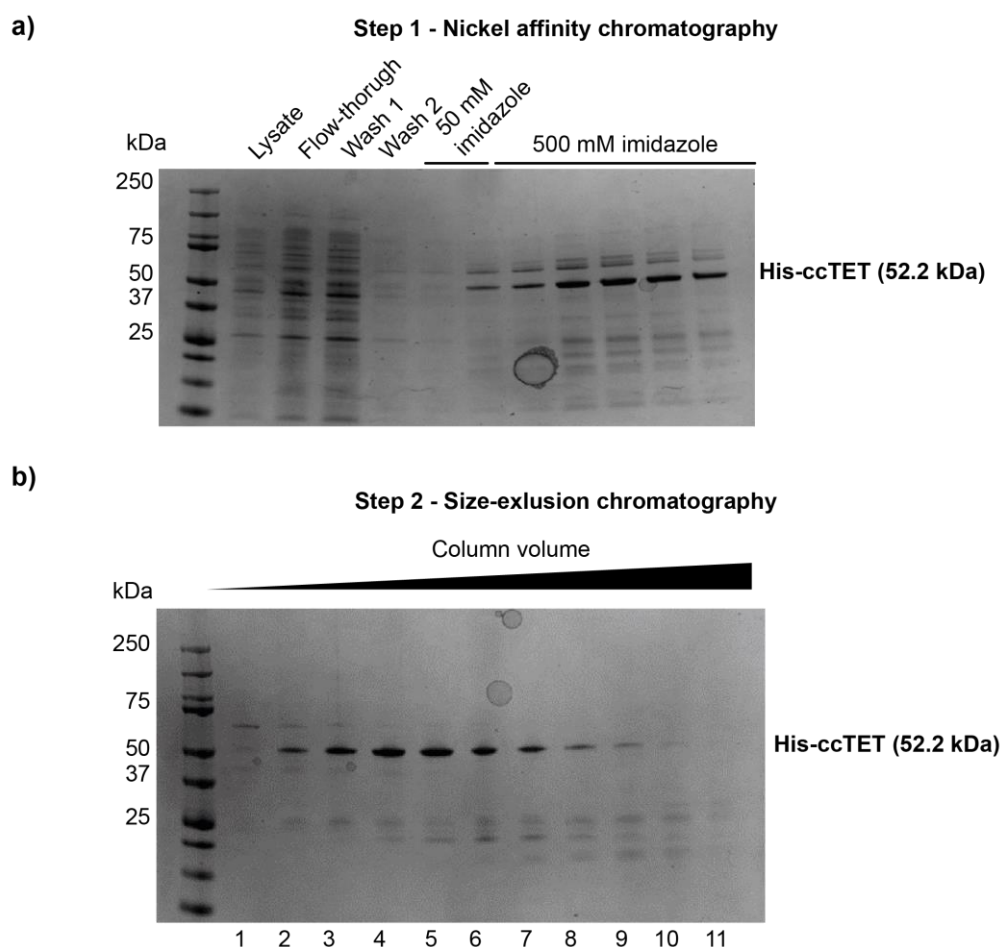


Figure 3-4 Representative, Coomassie-stained SDS PAGE electrophoresis gels for purification of His-ccTET. (a) SDS-PAGE gel showing protein fractions after His-ccTET purification on Ni-NTA resins. Flow-through contained unbound bacteria proteins. Wash 1 and wash 2 with low imidazole concentration removed majority of contaminant proteins. His-tagged ccTET protein was then eluted after washing with the buffers containing 50 mM and 500 mM imidazole. All eluted fractions were used for size-exclusion chromatography. (b) SDS PAGE after second step of His-ccTET purification by size-exclusion chromatography. Fractions from lanes 2-9 were collected, concentrated and used in further assays.

3.5 5mC in cfDNA can be enriched by chemical labelling

To ensure specific 5mC enrichment, hydroxymethylated sites had to be first efficiently blocked with UDP-glucose. We were wondering if it would be possible to perform 5hmC blocking and 5mC labelling in a single reaction by using a mixture of UDP-glucose and azide-modified glucose during the 5mC labelling step. This would allow to avoid a purification step between 5hmC blocking and 5mC labelling, reduce DNA losses and

make the method more suitable for low input applications. We first used Csp6I restriction digestion assay to test ability of UDP-glucose to block 5hmC. We used model DNA containing single 5hmC inside the Csp6I restriction site and subjected it to enzymatic glucosylation by T4-BGT enzyme using different concentrations of UDP-glucose (0.5-10 μ M). Subsequently, glucosylated products were digested by Csp6I enzyme and visualised on agarose gel. Without 5hmC protection, model DNA was digested into two short $\sim$ 50 bp fragments but after glucosylation enzymatic cleavage was blocked and 100 bp DNA remained intact (Figure 3-5, a). This assay revealed that 5hmC can be efficiently protected even by very low concentration of UDP-glucose equal to 0.5 μ M (Figure 3-5, b).

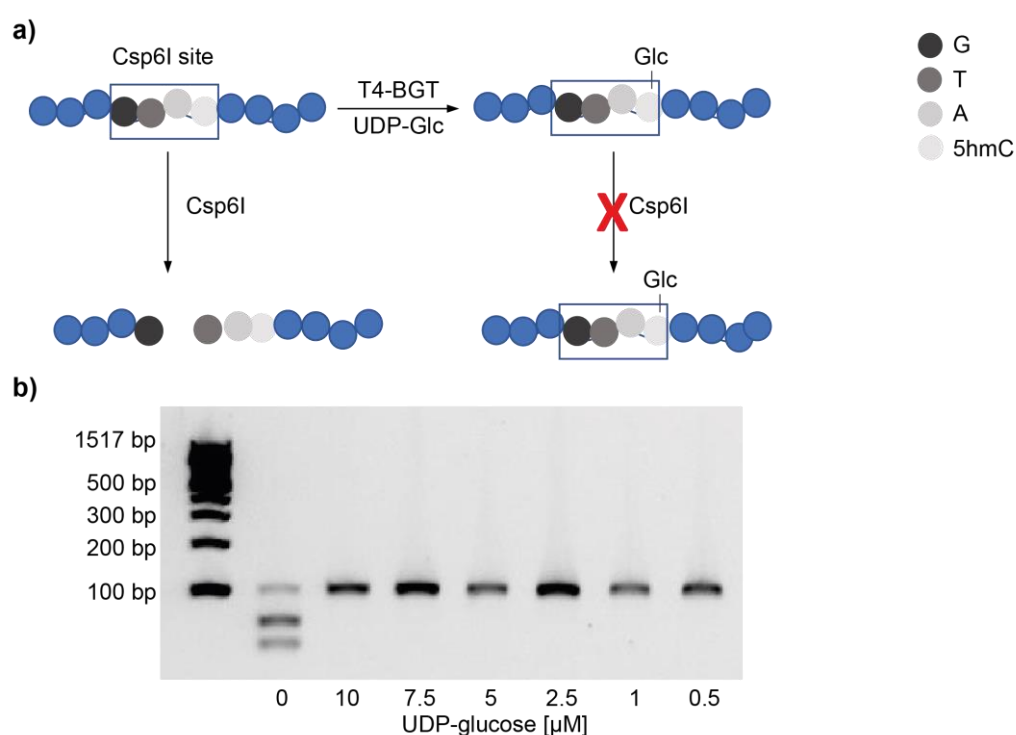


Figure 3-5. (a) Overview of Csp6I digestion assay for testing 5hmC protection. (b) Agarose gel of hydroxylated model DNA without UDP-glucose protection or protected with various concentration of UDP-glucose after Csp6I digestion. Without protection, model DNA was digested into two short fragments $\sim$ 50 bp. Incomplete digestion of unprotected DNA was probably caused by too short incubation time or not optimal conditions for Csp6I reaction. After glucosylation, model DNA was protected from Csp6I cleavage and remained intact. Concentration of model DNA in a blocking reaction was $\sim$ 30 nM. Even 0.5 μ M UDP-glucose efficiently protected tested DNA from digestion. Lower concentrations of UDP-glucose for 5hmC blocking were not tested.

As concentration of UDP-glucose required for effective 5hmC blocking was very low, we then investigated a one-pot blocking/labelling reaction protocol. To test feasibility of this approach, we first checked whether azide-modified glucose can be efficiently transferred to 5hmC when UDP-glucose is present in the reaction. Therefore, we designed Cy5.5 labelling assay where 5hmC-containing oligo was used for T4-BGT reaction with azide-modified glucose alone or with a mixture of azide-modified glucose and UDP-glucose. This reaction was followed by the transfer of Cy5.5 in a click chemistry. Labelling could be then visualised on 2% agarose gel (Figure 3-6, a). Although azide-modified glucose was successfully transferred on 5hmC, even 1% amount of UDP-glucose abolished UDP-N₃-glucose transfer and completely blocked labelling reaction (Figure 3-6, b). UDP-glucose is a native substrate for T4-BGT with a high affinity to the enzyme, thus it successfully competes with a modified substrate even at low concentration and completely blocks the transfer of UDP-N₃-glucose. For this reason, 5hmC blocking step and 5mC labelling have to be separated into two individual reactions.

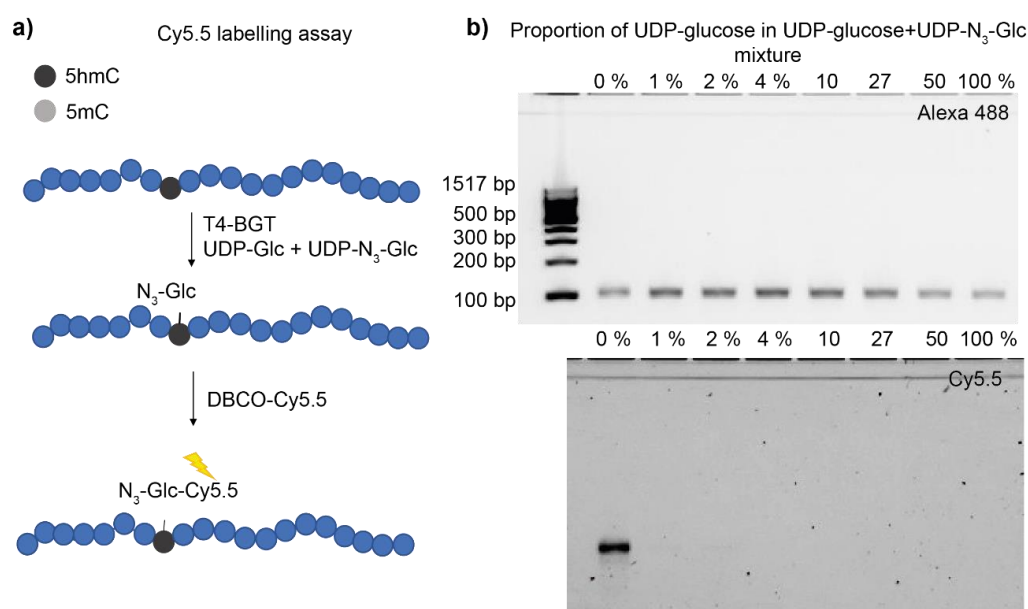


Figure 3-6. (a) Overview of Cy5.5 labelling assay for one-pot blocking/labelling reaction testing (b) Agarose gels of model DNA labelled with UDP-N₃-glucose or mixture of UDP-N₃-glucose and UDP-glucose. Top gel was visualised with Alexa 488 protocol that enables DNA visualisation. Bottom figure is the same gel visualised by Cy5.5 protocol that enables visualisation of labelled DNA. Only DNA glycosylated in the absence of UDP-glucose was

successfully labelled and visualised on the gel. Even 1% of UDP-glucose in the reaction completely blocked transfer of azide-modified glucose and abolished following Cy5.5 labelling.

We then performed similar Cy5.5 labelling assay to compare NgTET1 and ccTET performance in 5mC labelling reaction. Briefly, we subjected methylated model DNA to oxidation/labelling reaction and labelled it with DBCO-Cy5.5 (Figure 3-7, a). Reaction products were visualised on the agarose gel. Three various concentration of NgTET1 and ccTET were compared in the test. Labelling with NgTET1 resulted in higher Cy5.5 signal suggesting that this enzyme enables more efficient 5mC labelling (3-7, b). Therefore, we decided to use NgTET1 in further method development.

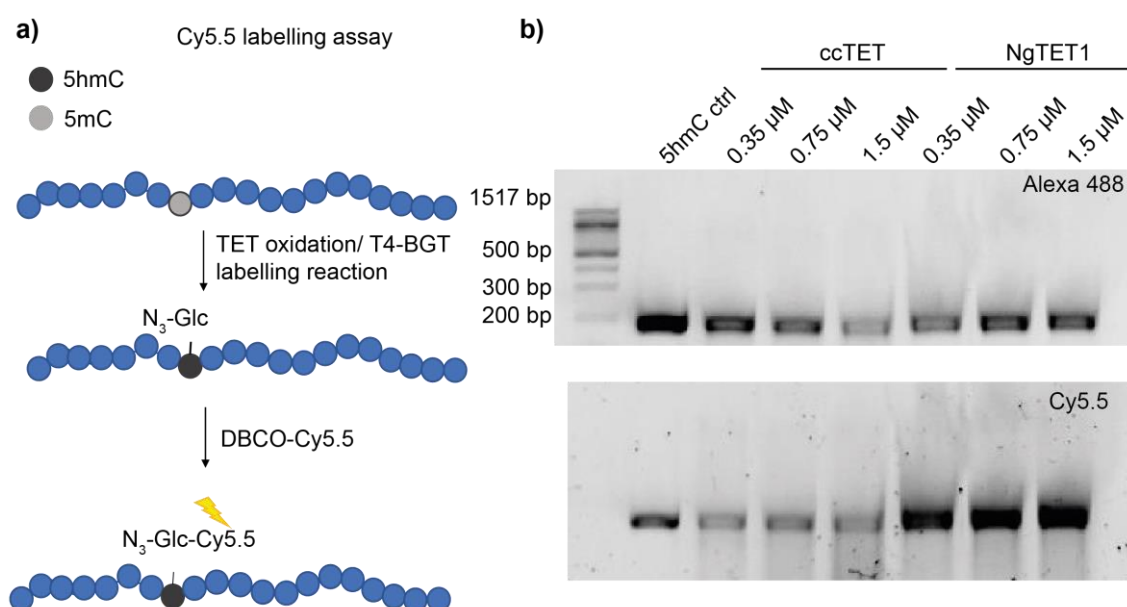


Figure 3-7. (a) Overview of Cy5.5 labelling assay for 5mC labelling reaction. (b) Agarose gels of methylated model DNA labelled in the reactions containing ccTET and NgTET1. 5hmC labelling reaction was used as a positive control. Top gel was visualised with Alexa 488 protocol that enables DNA visualisation. Bottom figure is the same gel visualised by Cy5.5 protocol that enables visualisation of labelled DNA. Labelling with NgTET1 at all tested concentrations was more efficient than ccTET. We assumed that better efficiency of 5mC labelling in reactions with NgTET1 was associated with higher oxidative activity of this enzyme. However, we cannot completely exclude the possibility that differences in purity of NgTET1 and ccTET and their subsequent effect on protein quantification affected these results.

Finally, we sought to apply the newly developed labelling reaction to cfDNA methylation enrichment for next generation sequencing. We adopted click-chemistry and pull-down conditions from the analogous method published for 5hmC enrichment (Song *et al.*, 2011,

2017). We tested different concentrations of NgTET1 in a labelling reaction to identify the best conditions for low-input DNA analysis. Efficiency and specificity of pull-down was tested with qPCR analysis of spike-in control oligos. Labelling reaction with low concentration of NgTET1 enabled specific enrichment of methylated fragments over unmodified or hydroxylated DNA. The highest fold enrichment was observed for control sequences with high density of CpG sites (Figure 3-8). Notably, the developed protocol was able to generate successful sequencing library with as low as 5 ng of cfDNA.

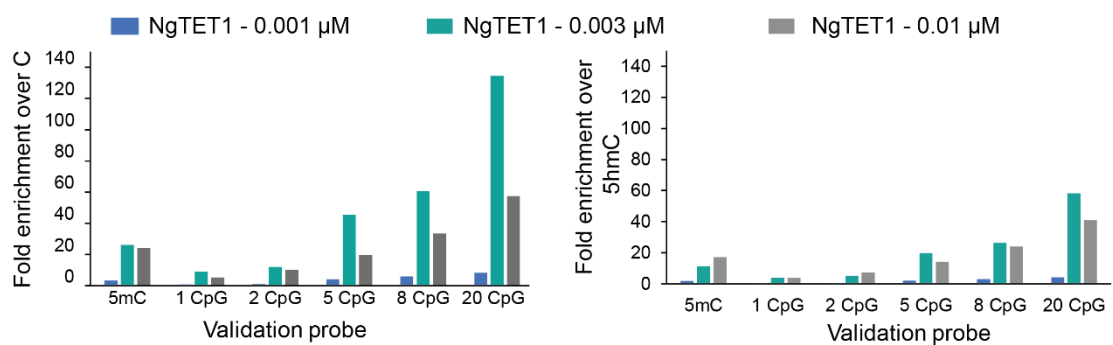


Figure 3-8. qPCR results of representative conditions for 5mC pull-down. To calculate fold enrichment, ΔC_t (cycle threshold) of particular spike-in control between input and enriched sample was calculated. Subsequently, fold enrichment value was calculated as ratio of ΔC_t of methylated and ΔC_t of unmodified spike-in controls (fold enrichment over C) or ΔC_t of methylated and ΔC_t of hydroxymethylated spike-in controls (fold enrichment over 5hmC). Experiment was performed once.

3.6 cfMeSeal sequencing provides specific methylation information

We applied newly developed cfMeSeal method to the cfDNA samples from 15 cancer patients and 8 healthy individuals to investigate the potential of this method for methylation profiling in clinical samples (Table 3-1). Sequencing quality was high and on average 14.6M (SD = 0.3M) of unique, mapped reads were generated per sample. Sequencing metrics are summarised in Appendix 3 Supplementary Table 3-1.

Table 3-1. Clinical characteristics of cfDNA samples used for cfMeSeal method

Sample ID	Type	Stage	Gender	Age
colon16	Colon cancer	II	Female	57
colon17	Colon cancer	IV	Male	52
colon19	Colon cancer	IV	Female	62
GBM57	Glioblastoma	IV	Female	52
GBM66	Glioblastoma	IV	Male	81
stomach2	Stomach cancer	III c	Male	54
stomach3	Stomach cancer	I a	Male	68
stomach8	Stomach cancer	I a	Male	65
pancreatic9	Pancreatic cancer	IV	Male	76
pancreatic15	Pancreatic cancer	IA	Male	64
pancreatic22	Pancreatic cancer	III	Female	71
pancreatic27	Pancreatic cancer	IV	Male	55
pancreatic68	Pancreatic cancer	IV	Male	63
pancreatic69	Pancreatic cancer	IIA	Male	66
pancreatic75	Pancreatic cancer	IIA	Male	54
H9	Healthy	NA	Male	56
H10	Healthy	NA	Male	52
H1083	Healthy	NA	Male	64
H2652	Healthy	NA	Female	66
H2654	Healthy	NA	Female	70
H2657	Healthy	NA	Female	65
H3455	Healthy	NA	Male	65
H3457	Healthy	NA	Female	51

Before analysis of cfDNA methylation profiles, efficiency and specificity of 5mC pull-down was assessed based on sequencing reads mapped to the spike-in validation probes. This analysis confirmed that cfMeSeal method specifically captured methylated cytosines with very low level of nonspecific enrichment of unmodified or hydroxymethylated sequences (Figure 3-9, a). As expected, we observed higher enrichment efficiency for control oligos containing more methylated CpG sites. On the other hand, initial quality control revealed differences in the enrichment patterns between two batches of the samples that were prepared in separate experiments. Although first batch exhibited an expected enrichment trend with the highest value observed for spike-in with the highest number of methylated CpG sites, enrichment profile in the second

batch was more irregular with more reads mapped to the moderately methylated spike-in controls (Figure 3-9, b).

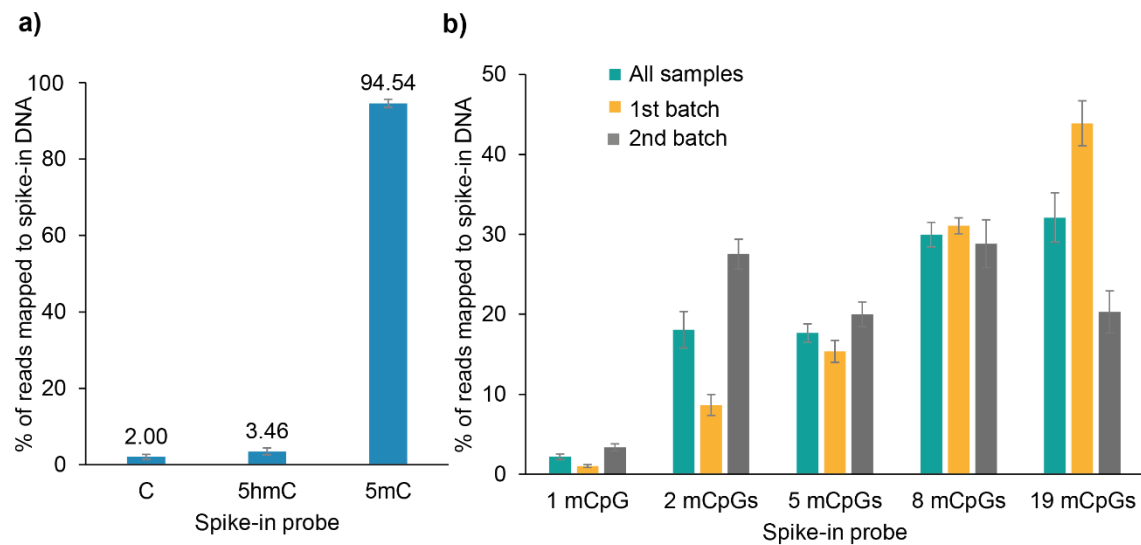


Figure 3-9. Percentage of sequencing reads mapped to the spike-in control oligos. Error bars represent standard error.

Existence of the batch effect in the dataset was further confirmed at cfDNA level. Unsupervised analysis of DNA methylation in 10 kb genomic windows showed clear batch-dependent separation between samples. This effect was attenuated after computational batch correction and all following results were generated on the batch-corrected data unless specified otherwise (Figure 3-10, a).

Notably, despite batch-dependent bias, there was good correlation between technical replicates of cfMeSeal libraries prepared from the same cfDNA in the separate experiments (Pearson correlation = 0.67, 0.87 and 0.93 for methylation in 1 kb, 5 kb and 10 kb genomic bins, respectively). Correlation was further improved after batch correction (Figure 3-10, b-c).

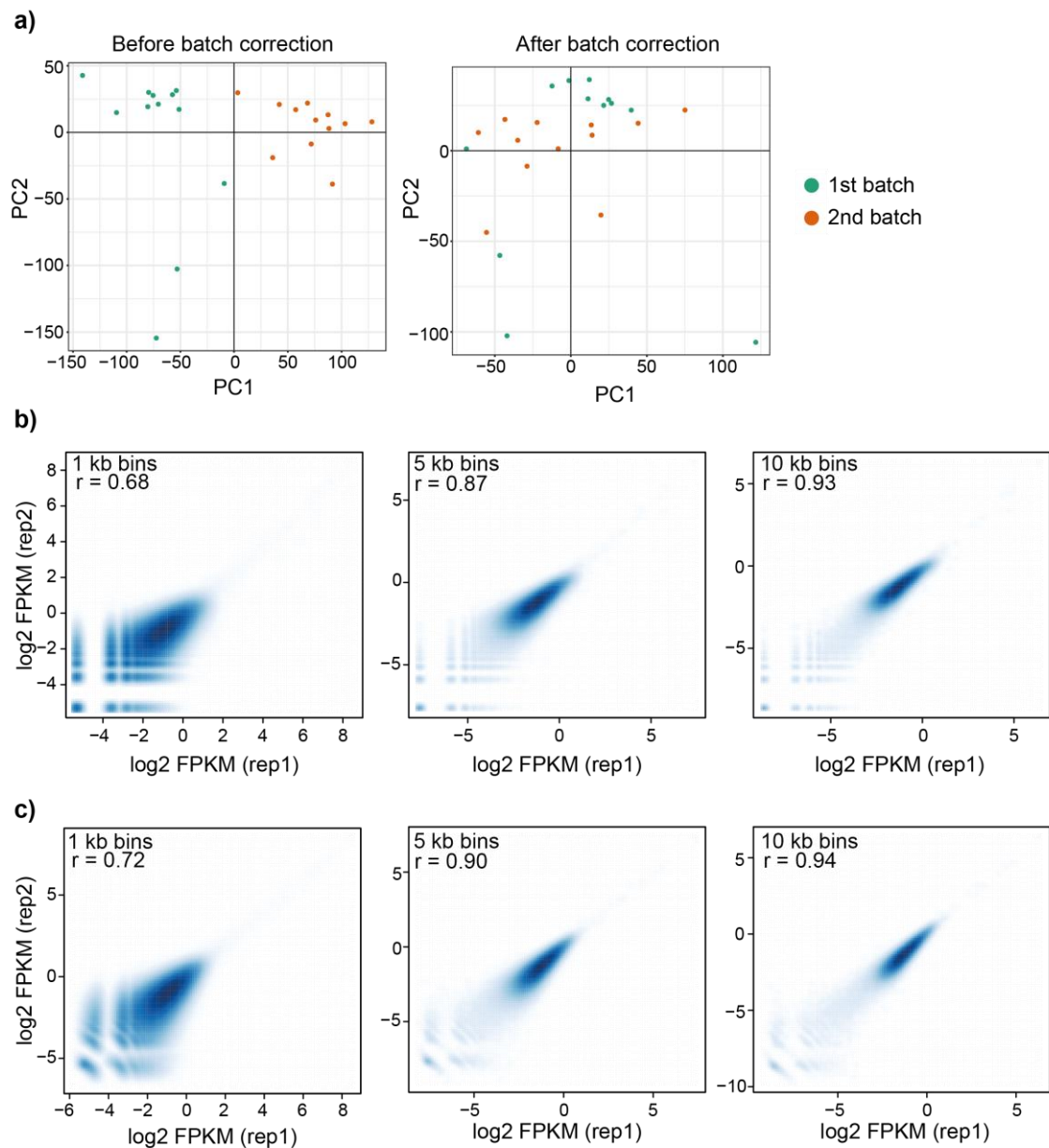


Figure 3-10. (a) Principal component analysis of cfDNA methylation in 10 kb genomic windows before and after batch correction. Batch correction was performed using *sva* R package (Leek *et al.*, 2014). (b) Correlation of 5mC log₂ FPKM values in 1 kb, 5 kb and 10 kb genomic windows between two technical replicates of cfMeSeal libraries. Pearson correlation coefficients are shown on the plots. Plots were prepared on data before batch correction. (c) Correlation of 5mC log₂ FPKM values in 1 kb, 5 kb and 10 kb genomic windows between two technical replicates of cfMeSeal libraries. Pearson correlation coefficients are shown on the plots. Plots were prepared on data after batch correction. Batch correction was performed using *sva* R package (Leek *et al.*, 2014).

3.7 cfMeSeal sequencing captures cancer-specific methylation patterns in cfDNA

After the satisfactory results of initial quality control, we sought to analyse cfDNA methylation profiles. We first looked at the cfMeSeal data tracks in Genome Browser and compared it with a track for historical whole-genome cfDNA sequencing prepared in our laboratory. We found that contrary to the whole genome sequencing that exhibited even coverage across the genome, cfMeSeal data had peaks and dips corresponding to the regions where methylation was enriched or depleted, respectively (Figure 3-11-a). 5mC detected by cfMeSeal showed characteristic genomic distribution already described in literature with methylation enriched in genebody and depleted from CGIs, TSS and TES (Lister *et al.*, 2009; Ficz *et al.*, 2011; Stadler *et al.*, 2011) (Figure 3-11, b-c). Metagene profiles of 5mC in genes ranked according to their expression based on public cfRNA-seq data (Song *et al.*, 2017) revealed that highly expressed genes exhibited lower methylation levels while lowly expressed and silenced genes were highly methylated (Figure 3-11, d) (Feng and Zhang, 2001; Borgel *et al.*, 2010; Maunakea *et al.*, 2010; Ficz *et al.*, 2011; Leung *et al.*, 2015). This together confirmed that cfMeSeal method was able to capture characteristic features of DNA methylation in the genome.

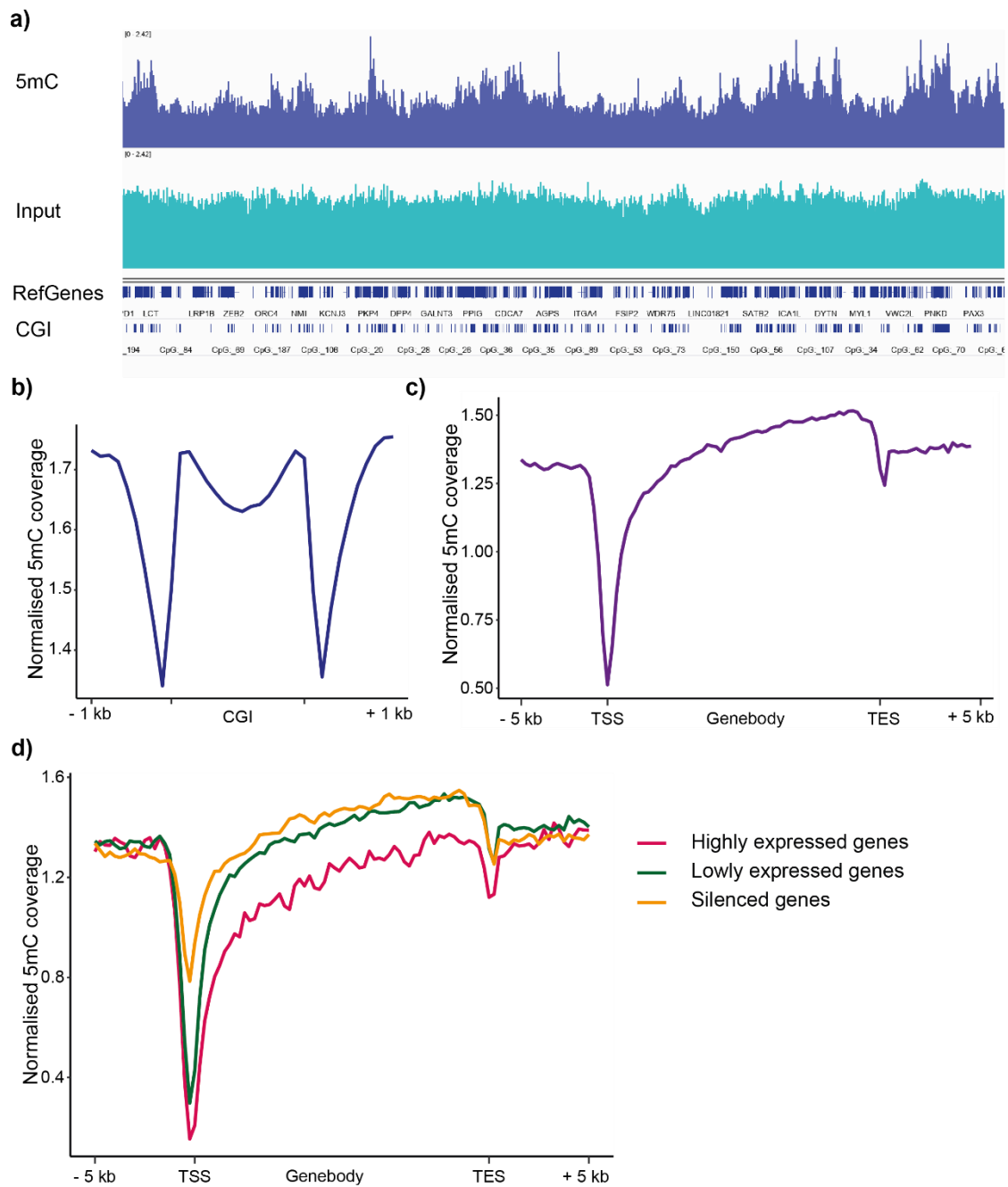


Figure 3-11. (a) Genome browser image of representative cfMeSeal data in comparison with historical whole-genome cfDNA sequencing data. (b) Metagene profile of cfDNA methylation across CGI and 4 kb flanking regions based on mean 5mC coverage in healthy controls. Normalised 5mC coverage was calculated by Deeptools (Ramírez *et al.*, 2014). (c) Metagene profile of cfDNA methylation across genebody and 5 kb flanking regions based on mean 5mC coverage in healthy controls cfDNA. Normalised 5mC coverage was calculated by Deeptools (Ramírez *et al.*, 2014). (d) Metagene profile of cfDNA methylation in representative cfDNA healthy control across genebody and 5 kb flanking regions in genes ranked according to their expression level. Normalised 5mC coverage was calculated by Deeptools (Ramírez *et al.*, 2014). Genes were ranked by their expression based on public cfRNA-seq data (Song *et al.*, 2017). Highly expressed genes were defined based on FPKM ≥ 10 . Lowly expressed genes were defined as genes with FPKM between 0.1 and 10. Silenced genes were defined as genes with FPKM ≤ 0.1 .

We then aimed to identify cancer-specific methylation patterns in cfDNA based on cfMeSeal sequencing. We found that pancreatic cancer samples separated from healthy in unsupervised analysis by PCA based on methylation in selected genomic features (Figure 3-12, a-d). Since the number of samples for other cancer types was too low to perform statistically significant differential methylation analysis, we decided to focus our further investigation on pancreatic cancer. We confirmed promising separation of most pancreatic cancer samples from healthy based on unsupervised analysis of methylation in promoter and genebody regions, which suggest cancer-specific methylation signatures in those samples (Figure 3-13, a-b).

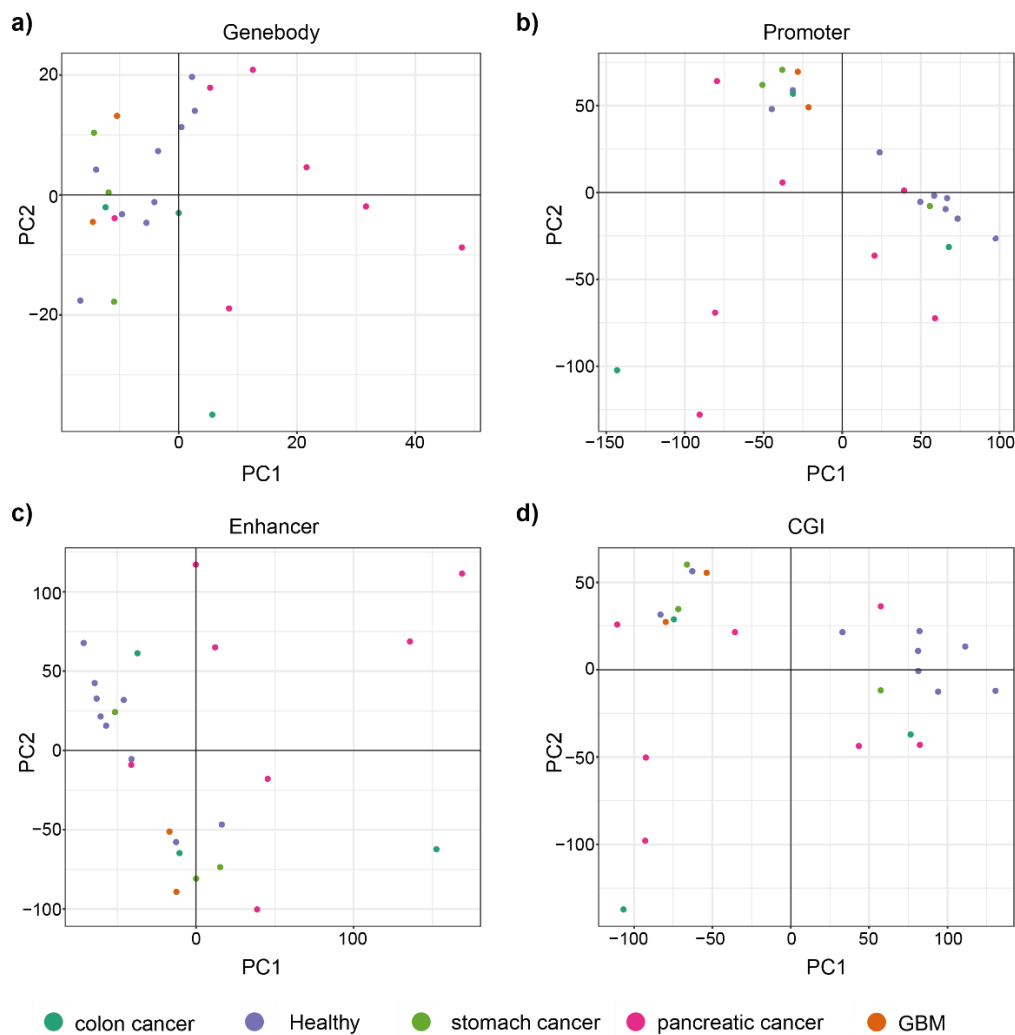


Figure 3-12. Principal component analysis of 5mC log₂ FPKM in (a) Genebodies, (b) gene promoters, (c) enhancers and (d) CGIs.

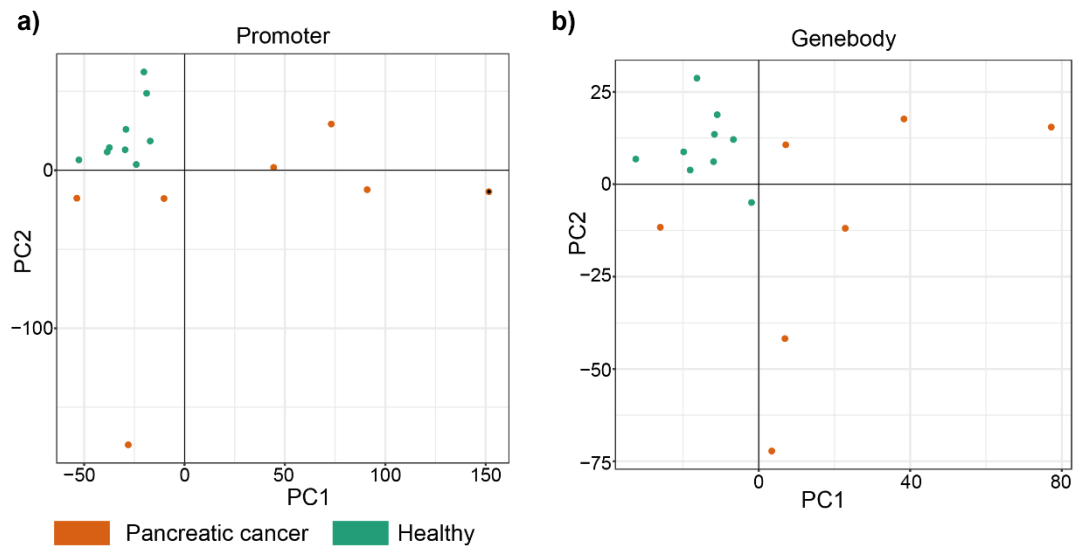


Figure 3-13. (a) PCA analysis of promoter 5mC log₂ FPKM for pancreatic cancer and healthy samples. (b) PCA analysis of genebody 5mC log₂ FPKM for pancreatic cancer and healthy samples.

Differential methylation analysis between plasma cfDNA from pancreatic cancer and healthy identified 3485 genes with significant differences in genebody methylation and 4101 differentially methylated promoters ($p < 0.05$; genebodies: average log₂ fold change: -0.12 and range: -6.40-6.27; promoters: average log₂ fold change: -0.36 and range: -2.32-2.40). Specifically, 1290 genebodies and 1358 promoters were hypermethylated in cfDNA from pancreatic cancer whereas 2195 and 2743 were hypomethylated (Figure 3-14, a-d). Differentially methylated genes included genes involved in metabolism (HK2, SLC2A2), tumour suppression (CDKN2A, TP53), cell growth (MTOR, IGF1), apoptosis (FAS, CASP8) and epigenetic regulation (DNMT1, KDM5A). Gene Ontology analysis revealed enrichment of differentially methylated regions in pathways associated with gland development, GTPase activity, metabolism, apoptosis and liver function (Figure 3-15, a-d)

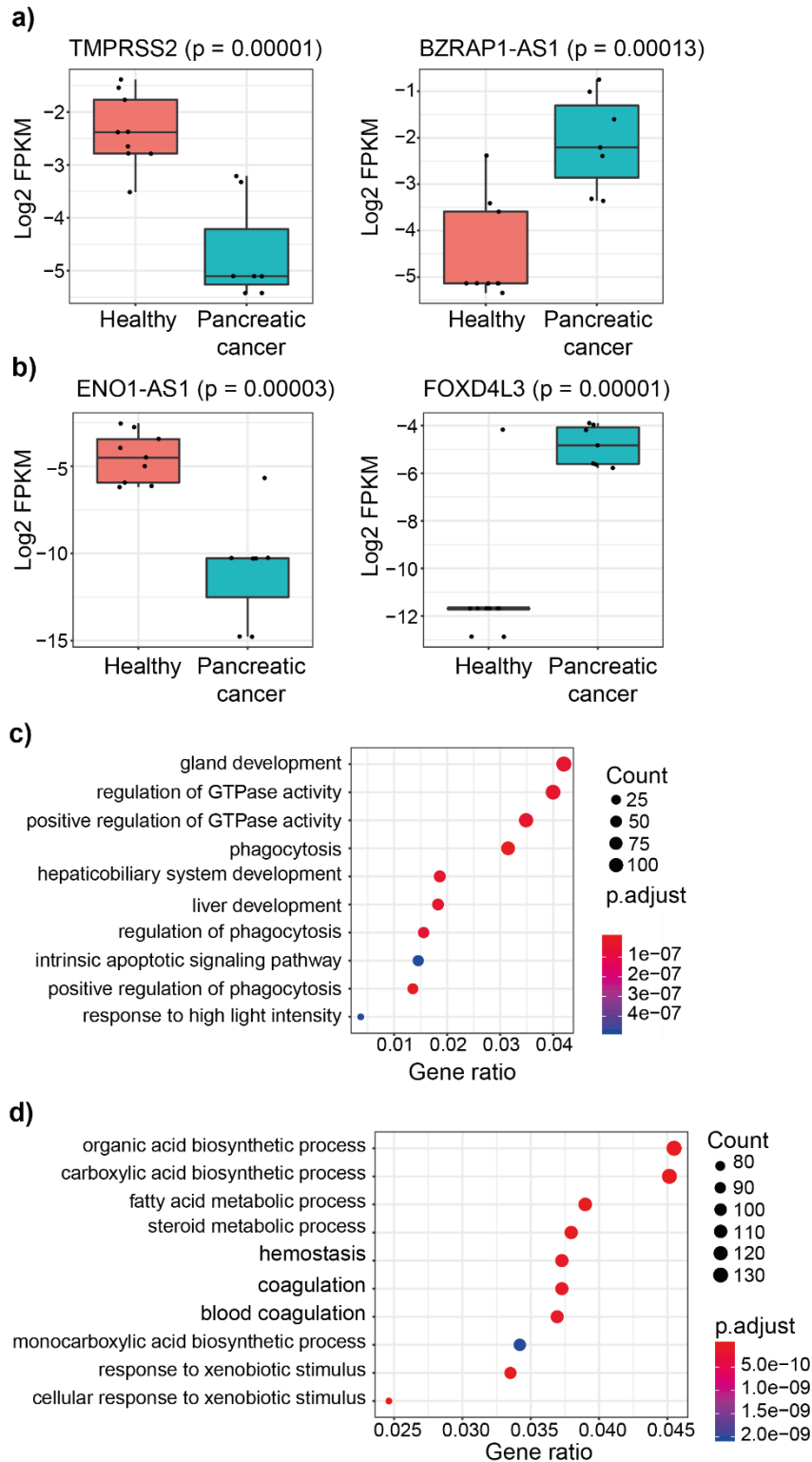


Figure 3-15. (a) Boxplots of top hyper- and hypomethylated gene promoters between pancreatic cancer and healthy. (b) Boxplots of top hyper- and hypomethylated gene bodies between pancreatic cancer and healthy. GO analysis of differentially methylated gene bodies (c) and gene promoters (d). Only 10 the most significant functional annotations are shown.

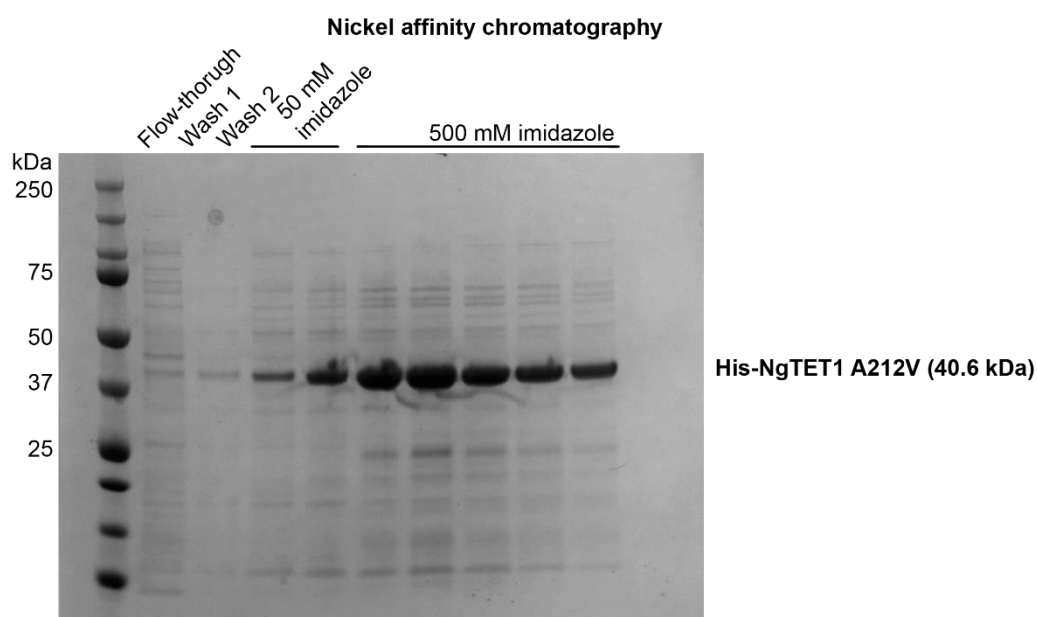
3.8 Mutated NgTET1 does not improve 5mC labelling reaction stability

The batch effect observed in first cfMeSeal sequencing could be potentially related to problems with reproducibility of labelling reaction. In fact, 5mC labelling is based on assumption that methylated sites after TET oxidation to 5hmC will be immediately labelled with azide-glucose before they can be further oxidised to 5fC and 5caC. This is based on the previous observations that 5hmC to 5fC oxidation kinetics is slower than T4-BGT reaction (Terragni *et al.*, 2012; Hu *et al.*, 2015; Pais *et al.*, 2015). However, it is possible that some factors can disturb this balance and lead to overoxidation of 5mC to 5fC and 5caC so it could not be labelled. Since high density CpG regions are more likely to be oxidised, such a situation could result in an overoxidation of spike-in controls with high CpG density and their underrepresentation in the pull-downed samples with simultaneous high enrichment of methylated CpG-sparse regions – situation similar to that observed in the second batch of sequenced cfDNA samples. The inhibition of oxidation reaction after generation of 5hmC could then help to improve stability and reproducibility of labelling.

Mutagenesis of TET proteins to decrease efficiency of 5hmC oxidation could help to achieve this goal. A mutated version of NgTET1 protein where alanine at position 212 is changed to valine (NgTET1 A212V) was described in literature. As reported, this mutant exhibited limited ability to oxidise 5hmC to 5fC and 5caC while it retained almost normal efficiency of 5mC oxidation to 5hmC (Hashimoto *et al.*, 2015). We performed mutagenesis of wild type NgTET1 plasmid and expressed and purified NgTET1 A212V protein (Figure 3-16, a-b). We tested its oxidation activity on short 11mer oligo with single 5mC site with MALDI MS. It revealed that contrary to the wild type NgTET1 that can completely oxidise 5mC to 5caC, the main product of NgTET1 A212V oxidation is

5hmC and/or 5fC. HPLC analysis of mESC genomic DNA oxidised with NgTET1 A212V showed that mutated protein was still able to produce a substantial proportion of 5fC and 5caC, however, lower than those generated by the wild type NgTET1 (Figure 3-17, a-c).

a)



b)

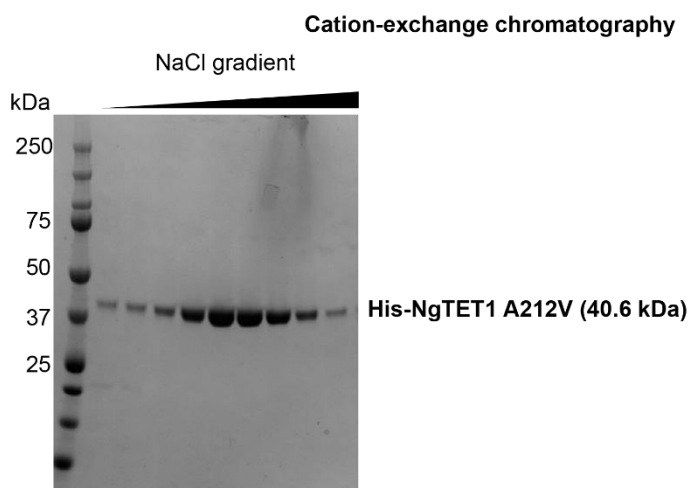


Figure 3-16. Representative, Coomassie-stained SDS PAGE electrophoresis gels for purification of His-NgTET1 A212V. (a) SDS-PAGE gel showing protein fractions after His-NgTET1 A212V purification on Ni-NTA resins. Flow-through contained unbound bacteria proteins. Wash 1 and wash 2 with low imidazole concentration removed majority of contaminant proteins. His-tagged NgTET1 A212V protein was then eluted after washing with buffers containing 50 mM and 500 mM imidazole. All eluted fractions were used for size-exclusion chromatography. SEC was performed and fractions were collected based on chromatograph and used for ion exchange chromatography without running SDS PAGE gel. (b) SDS PAGE gel of pure His-NgTET1

A212V protein fractions after cation-exchange chromatography. All protein-containing fractions were collected, concentrated and used in further assays.

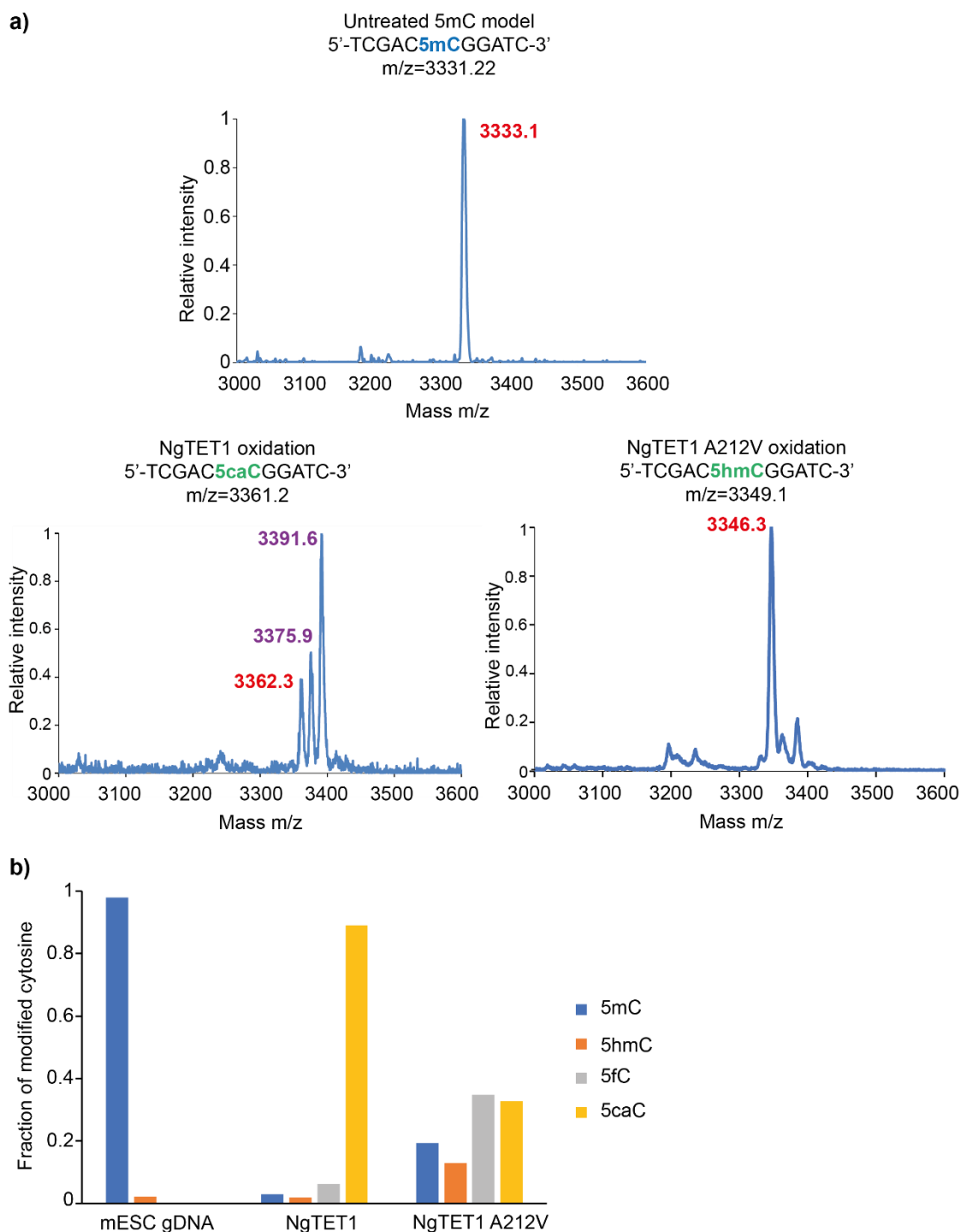


Figure 3-17. Representative results of TET activity assay measured by MALDI. MALDI characterization of 5 μ M 11mer 5mC DNA model before and after oxidation with 5 μ M NgTET1 or NgTET1 A212V. Calculated mass of 11mer after complete 5mC to 5caC oxidation was shown in red. Additional peaks after oxidation with NgTET1 shown in purple are results of off-target T oxidation to fdU (3375.9) and cadU (3391.6). Peak at m/z of 3346.3 after NgTET1 A212V can correspond both to 5hmC and 5fC that cannot be confidently distinguished on MALDI due to too small difference in molecular weights. (b)

HPLC–MS/MS quantification of relative modification levels in the mESCs genomic DNA control, after oxidation by NgTET1 or NgTET1 A212V. Experiment was performed once.

Although NgTET1 A212V exhibited higher oxidation efficiency than expected, its activity on 5hmC was still lower compared to the wild-type enzyme and thus it could still be useful for cfMeSeal development. We performed optimisation of 5mC labelling reaction with various concentrations of NgTET1 A212V and T4-BGT. NgTET1 A212V allowed for specific and efficient labelling of 5mC based on end-point PCR assay, however, enrichment profiles were still poorly reproducible between replicates of the same conditions validated by qPCR suggesting that mutated NgTET1 is not able to improve stability and reproducibility of cfMeSeal method compared to the wild-type NgTET1 (Figure 3-18, a-b).

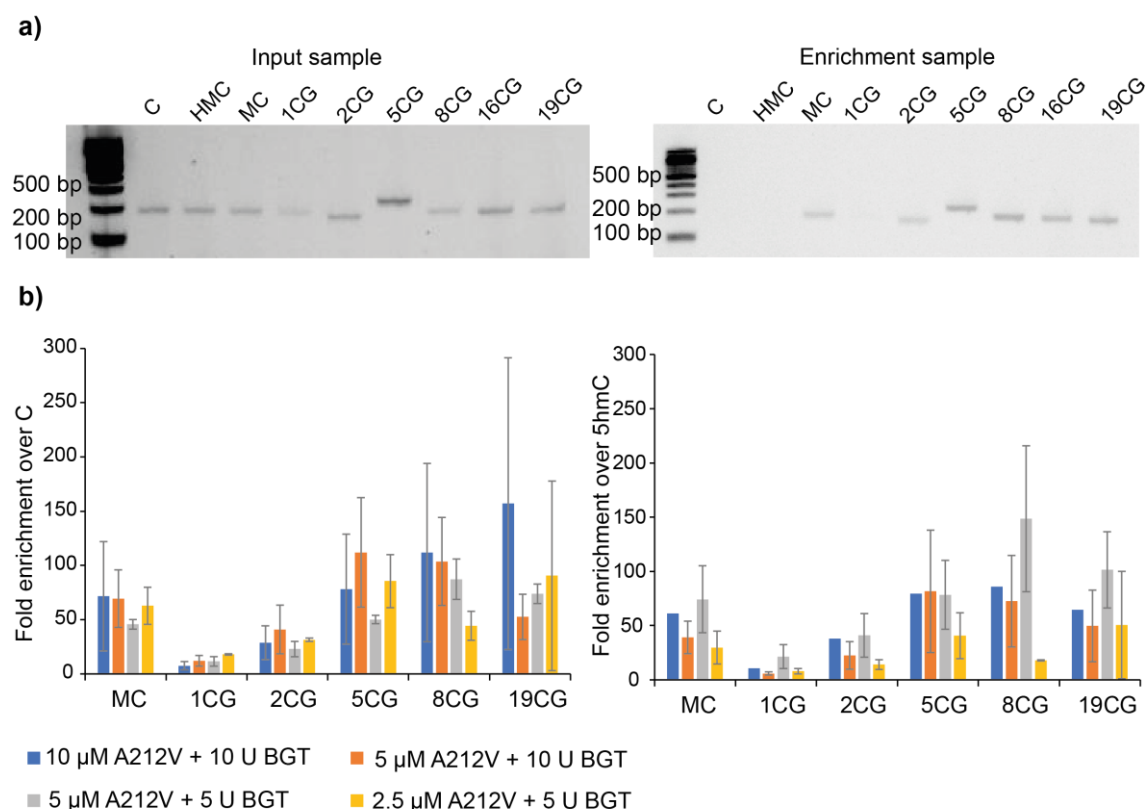


Figure 3-18. (a) Representative agarose gel of end-point PCR of spike-in control DNA in cfDNA sample before and after 5mC pull-down by cfMeSeal method. (b) qPCR results of 5mC pull-down optimisation with NgTET1 A212V. To calculate fold enrichment, ΔC_t of particular spike-in control between input and enriched sample was calculated. Subsequently, fold enrichment value was calculated as ratio of ΔC_t of methylated and ΔC_t of unmodified spike-in controls (fold

enrichment over C) or ΔCt of methylated and ΔCt of hydroxymethylated spike-in controls (fold enrichment over 5hmC). Error bars represent standard error.

3.9 Discussion

In this work, we developed a novel cfMeSeal method and applied it for cfDNA sequencing. We demonstrated that our method is able to specifically enrich for methylated cfDNA fragments and provides genome-wide methylation even from 5 ng of sample. In comparison to the existing technologies, this approach is mild and non-damaging and does not rely on antibody-based enrichment.

3.9.1 cfDNA methylation patterns in cancer

We confirmed that cfMeSeal sequencing was able to capture characteristic genomic distribution of 5mC including depletion in CGI and highly transcribed genes and enrichment in gene bodies and silenced genes. We also found that even in unsupervised analysis we were able to find cancer-specific methylation patterns in the important genomic regions including genebodies, promoters, enhancers and CGI. PCA analysis of all cancer types and healthy showed that cancer samples did not form common, separate cluster, which suggests that cfDNA methylation profiles are heterogenous between cancer types and therefore merging them as a single cancer group could not be justified. Due to insufficient number of samples for most cancer types included in this sequencing we did not perform differential methylation analysis for all of them. Instead, we decided to focus on pancreatic cancer since it exhibited the most promising separation from healthy on PCA plot and had the biggest sample size among the cancer types. We found 3485 genes with significant differences in genebody methylation and 4101 differentially methylated promoters. Functional analysis determined that these genes are enriched in pathways related to metabolism of fatty acids and steroids as well as liver and

hepatobiliary system development, which corresponds to the pancreatic function as well as to liver function. To note, liver function is commonly affected in pancreatic cancer patients (Neoptolemos *et al.*, 2018). As most of cfDNA is derived from blood cells, GO analysis identified enrichment of differentially methylated regions in pathways related to haemostasis (S. Guo *et al.*, 2017). This together suggest that aberrant methylation signal detected in plasma cfDNA from pancreatic cancer patients is likely to be a mixture of cancer-specific methylation patterns derived both from tumour and somatic cells (e.g. blood cells). Interestingly, our analysis was able to capture known pancreatic cancer epigenetic changes in cfDNA including hypermethylation of promoters for tumour suppressor genes such as TP53 and CDKN2A as well as hypomethylation of KRAS promoter (Tang *et al.*, 2015; Natale *et al.*, 2019). These observations suggest that even with shallow sequencing we were still able to detect methylation marks likely derived from tumour cells. Differentially methylated genes also included numerous epigenetic modulators such as hypomethylation of DNMT3B and histone demethylases in line with previous reports about aberrant epigenetic landscape in pancreatic cancer (Lomberk *et al.*, 2018).

Top differentially methylated genes identified by our analysis have distinct functions and include transmembrane proteins (TMPRSS2, OR4C11) (Malnic, Godfrey and Buck, 2004; Lam *et al.*, 2015), gene expression regulators (MYNN, FOXD4L3) (Alliel *et al.*, 2000; Ramakrishna *et al.*, 2010), cell death (FAS, SPNS1) (Xiang and Tong, 2008; Peter *et al.*, 2015), lncRNA (BZRAP1-AS1, ENO1-AS, BRE-AS) (Li *et al.*, 2018; M. Zhang *et al.*, 2018), metabolic processes (PC) (Xiang and Tong, 2008; Monnot *et al.*, 2009) and immune system function (SIGLEC14, SCGB3A1) (Krop *et al.*, 2001; Guo *et al.*, 2008; Graustein *et al.*, 2017; Jimmy Huang *et al.*, 2018). Importantly, dysregulation of some of

those genes has been already identified in pancreatic cancer and other human malignancies.

These results confirmed that enrichment-based methylation sequencing can provide important information about cfDNA methylation. This is in good agreement with recently published cfMeDIP papers that also proved that cancer-specific methylation changes can be detected in cfDNA and that cfDNA methylation has predictive potential (Shen *et al.*, 2018; Nassiri *et al.*, 2020; Nuzzo *et al.*, 2020). Unfortunately, as our study was a pilot experiment, the number of samples was not sufficient to build a reliable classifier and directly compare cfMeSeal predictive potential with cfMeDIP.

It is important to note, that design of our study had some limitations that could influence differential methylation results. First, batch effect and computational batch correction could affect the raw data and influence subsequent downstream analysis. Second, small sample size, different sources of collected plasma and cfDNA samples, as well as other potential confounding factors not considered during data analysis (e.g. age, diet, comorbidities, BMI, alcohol consumptions or smoking status), could contribute to identified DMRs. Therefore, it is important to anticipate that these results can be partially affected by other factors rather than biological mechanisms and would require further experimental validation. On the other hand, we were able to detect cancer-related methylation features already known in the literature and related to pancreatic cancer, which suggests that our data still provided biologically relevant information.

To summarise, cfMeSeal sequencing was able to *de novo* identify differentially methylated regions in cfDNA from cancer patients. Detection of aberrant methylation in previously reported cancer-related genes as well as genes related to immune cells and blood cell function suggest that both tumour and blood cells were potential source of observed methylation changes in pancreatic cancer. Those features could be investigated

as potential biomarkers and their ability to detect cancer may be validated on a bigger, independent cohort of patients.

3.9.2 Limitations

One limitation to the current cfMeSeal protocol is the batch effect. This issue should be particularly addressed before the method is applied to clinical samples in the future. We initially attributed lack of reproducibility to 5mC oxidation/labelling reaction as we hypothesised that overoxidation of 5mC to 5fC/5caC could have caused enrichment problems. We tried to optimise cfMeSeal method with mutated enzyme, NgTET1 A212V, with weaker activity on 5hmC, however, it did not improve overall stability and reproducibility of the method.

There are few possible explanations for these observations. First, NgTET1 A212V still retains ability to oxidise 5hmC to 5fC and 5caC, although with reduced rate. This decrease in activity might have not be enough to prevent from overoxidation. To note, NgTET1 A212V 5hmC oxidation efficiency observed by us is higher than reported in literature (Hashimoto *et al.*, 2015). Potential contamination with wild- type NgTET1 or mutagenesis failure were excluded as plausible reasons for this observation because identity of mutated NgTET1 was confirmed by TOF-MS (Appendix 3 Supplementary Figure 3-1).

Another source of variable enrichment patterns by cfMeSeal could be off-target activity of NgTET1 to oxidise thymine to uracil derivatives (Pais *et al.*, 2015). Although this reaction is thought to occur at the very low level in genomic DNA, we did not quantify it and we cannot rule out its potential impact on low input samples. Since we did not observe high level of unspecific pull-down of unmodified or hydroxymethylated spike-in controls, it is unlikely that thymine off-target oxidation to 5-hydroxymethyluracil by

NgTET1 and its unspecific labelling with azide-glucose contributed to the enrichment of cfDNA fragments regardless of their methylation status and was responsible for observed reproducibility problems.

Another source of uneven enrichment in cfMeSeal samples could be adapter ligation bias. In our protocol, sequencing adapters are ligated after labelling. Although this strategy improved library yield, heavily biotinylated high CpG density regions could be less efficiently ligated to the adapters resulting in an uneven enrichment. Finally, highly biotinylated fragments may not be efficiently amplified by the standard, commercially available DNA polymerases used in this study (Taq polymerase and KAPA HiFi HotStart DNA Polymerase) resulting in their underrepresentation in the final library. However, it is difficult to explain why ligation or PCR bias was observed only in some sample batches. To improve cfMeSeal method in the future, one would need to determine whether source of enrichment irreproducibility is labelling reaction, library preparation or amplification bias. Quantitative PCR could help to determine whether highly labelled fragments are amplified with lower efficiency, however, this approach can be affected by PCR artefacts and polymerase-specific biases. Therefore, careful experimental design including technical replicates and various DNA polymerases could help more reliable assessment of post-labelling amplification efficiency as well as find optimal conditions for an efficient amplification of labelled DNA fragments. Another direction to improve cfMeSeal method is replacing NgTET1 enzyme with a mammalian counterpart with lower off-target activity on thymine, e.g. mTet1CD. This could help to improve enrichment specificity if the problem is indeed caused by non-target oxidation. Moreover, mammalian TET enzymes do not interfere with single-stranded adapter DNA, therefore, it would be possible to perform adapter ligation step prior to the labelling reaction to eliminate risk of ligation bias. Further mutagenesis of mammalian TET enzymes to

reduce their ability to oxidise 5hmC could be also applied. Potentially useful mutants have been already proposed for human Tet2 enzyme (Liu *et al.*, 2017). Alternatively, careful structural analysis could help to design new mutagenesis sites and be a starting point for generating new TET enzymes with modified activities.

Finally, it is important to note that the experimental design for this project did not allow for tight control of the batch effect. Many experiments were prepared just once and technical replicates were not routinely included at the optimisation stage. Therefore, reproducibility of cfMeSeal method was not accurately assessed before sequencing. Some method conditions including reaction incubation times and temperatures as well as set of spike-in DNA controls were kept constant during experiments. However, home-made reaction buffers and changes in NgTET1 activity due to the different enzyme batch and long-term storage could have contributed to the observed reproducibility problems. Careful control of reagents, enzymes and reaction conditions during cfMeSeal as well as technical replicates in all the experiments could help to decrease variability in 5mC enrichment and enable more reliable assessment of method reproducibility before its application to the clinical samples and costly sequencing experiments.

3.9.3 Concluding remarks

In this part of the project, we modified and optimised a 5mC labelling and enrichment method for cfDNA methylation sequencing. This protocol was able to specifically enrich methylated fragments in cfDNA and generate successful sequencing libraries from as low as 5 ng of cfDNA. We demonstrated that this method can capture characteristic methylation distribution in the genome and is able to identify potential cancer-specific methylation patterns in cfDNA.

3.10 Materials and Methods

3.10.1 Preparation of spike-in control oligos

Spike-ins control oligos were designed as 160-180 bp DNA fragments containing different numbers of unmodified cytosines, hydroxymethylated cytosines or different number of methylated CpG sites. Control oligos were first amplified in PCR reaction containing 1 ng lambda DNA as template (Promega, D1501), 0.2 μ M primers, 1.25 U Taq DNA Polymerase and 200 μ M dNTPs (standard cocktail: dATP/dGTP/dTTP/dCTP excluding HMC probe: dATP/dGTP/dTTP/90%dCTP+10% dhmCTP and MC probe - dATP/dGTP/dTTP/90%dCTP+10% dmCTP). Amplicons were purified on Zymo-Spin IC Columns according to manufacturer's protocol.

To produce methylated control spike-ins probes CpG_1, CpG_2, CpG_5, CpG_8, CpG_16 and CpG_19 were CpG-methylated in 20 μ L reaction containing 3.2 U M. SssI (NEB), 0.64 mM SAM in NEB2 Buffer according to manufacturer's protocol (1h at 37°C). Amplicons were purified on Zymo-Spin IC Columns according to manufacturer's protocol.

All purified spike-in oligos were stored at -20°C. Sequences of control oligos and PCR primers are listed in Appendix 3 Supplementary Table 3-2.

3.10.2 Expression and purification of NgTET1

Details of recombinant NgTET1 are in Appendix 3 Supplementary Table 3-3. The pRSET-A plasmid encoding His-tagged NgTET1 protein (GG739552.1) was designed and purchased from Invitrogen. Protein was expressed in *E. coli* BL21 (DE3) bacteria and purified as previously described with some modifications (Hashimoto et al., 2014; Pais et al., 2015). Briefly, for protein expression, bacteria from overnight small-scale

culture were grown in LB medium at 37°C and 200 r.p.m. until the absorbance A₆₀₀ was between 0.7 and 0.8. Then cultures were cooled down to room temperature and target protein expression was induced with 0.2 mM isopropyl- β -D-1-thiogalactopyranoside. Cells were maintained for an additional 18 h at 18°C and 180 r.p.m. Subsequently, cells were harvested and re-suspended in the buffer containing 20 mM HEPES (pH 7.5), 500 mM NaCl, 1 mM dithiothreitol, 20 mM imidazole, 1 μ g ml⁻¹ leupeptin, 1 μ g ml⁻¹ pepstatin A and 1 mM PMSF. Cells were broken with EmulsiFlex-C5 high-pressure homogenizer and lysate was clarified by centrifugation for 1 h at 30,000g and 4°C. Collected supernatant was loaded on Ni-NTA resins and NgTET1 protein was eluted with buffer containing 20 mM HEPES (pH 7.5), 500 mM imidazole, 2 M NaCl, 1 mM dithiothreitol. Collected fractions were then purified on HiLoad 16/60 Sdx 75 (20 mM HEPES pH 7.5, 2 M NaCl, 1 mM dithiothreitol). Fractions containing NgTET1 were then collected, buffer exchanged to the buffer containing 20 mM HEPES (pH 7.0), 10 mM NaCl, 1 mM dithiothreitol and loaded on HiTrap HP SP column. Pure protein was eluted with the salt gradient, collected and buffer exchanged to the final buffer containing 20 mM Tris-Cl (pH 8.0), 150 mM NaCl and 1 mM dithiothreitol. Protein was then concentrated up to 130 μ M, mixed with glycerol (30% v/v) and aliquots were stored at -80°C.

3.10.3 Expression and purification of ccTET

Details of recombinant ccTET are in Appendix 3 Supplementary Table 3-4. The pRSET-A plasmid encoding His-tagged ccTET protein (CC1G_05589) was designed and purchased from Invitrogen (Zhang *et al.*, 2014). Protein was expressed in *E. coli* BL21 (DE3) bacteria and purified by Nickel-affinity and size-exclusion chromatography as described above for NgTET1. After SEC, fractions containing protein were collected,

concentrated and buffer exchanged to the final buffer containing 20 mM Tris–Cl (pH 8.0), 150 mM NaCl and 1 mM dithiothreitol. Protein was mixed with glycerol (30% v/v) and aliquots were stored at –80°C.

3.10.4 Preparation of NgTET1 A212V

Details of recombinant NgTET1 A212V are in Appendix 3 Supplementary Table 3-5. NgTET1 A212V plasmid was prepared by site-specific mutagenesis of wild-type pRSET-A NgTET1 plasmid. Mutagenesis was performed by Q5®Site-Directed Mutagenesis Kit (NEB) according to manufacturer's protocol with following primers: Forward – 5'-TCCGAGCACCGtgTTTAATAGCC-3' and Reverse 5'-TACAGACAATTATTCAGGG-3'. Mutagenesis was validated based on Sanger sequencing.

NgTET1 A212V protein was expressed and purified as described above for NgTET1.

3.10.5 TET activity test

For NgTET1 and NgTET1 A212V activity assay 5 µM 5mC 11mer oligonucleotide was oxidised in 20 µL reaction containing 50 mM MOPS buffer (pH 6.9), 100 µM ammonium iron(II) sulfate, 1 mM α-ketoglutarate, 2 mM ascorbic acid, 1 mM dithiothreitol, 50 mM NaCl and 5 µM NgTET1 for 1 h at 37 °C. After that, 4 U of Proteinase K (New England Biolabs) were added to the reaction mixture and incubated for 30 min at 37°C.

To test enzyme activity on genomic DNA, mESC DNA was incubated in 50 µl solution containing 50 mM MOPS buffer (pH 6.9), 100 µM ammonium iron(II) sulfate, 1 mM α-ketoglutarate, 2 mM ascorbic acid, 1 mM dithiothreitol, 50 mM NaCl and 5 µM NgTET1 for 1 h at 37 °C. After that, 4 U of Proteinase K (New England Biolabs) were added to the reaction mixture and incubated for 30 min at 37°C. The product was cleaned up on

1.8× AMPure XP beads following the manufacturer's instruction. Oxidised DNA was analysed by HPLC-MS/MS.

3.10.6 Optimisation of UDP-glucose concentration for 5hmC blocking

Minimal concentration of UDP-glucose required for 5hmC blocking was determined based on 100 bp model DNA containing hydroxymethylated cytosine in Csp6I restriction site. Sequence of the oligo is in Appendix 3 Supplementary Table 3-6. 20 ng of model DNA was blocked in 10 µL reaction containing 50 mM HEPES pH = 8.0, 25 mM MgCl₂, 1 mM DTT, 1 µL T4-βGT (ThermoFisher) and 0-10 µM UDP-glucose (Zymo) at 37°C for 1 h. Then reactions were purified on Zymo-Spin column (Zymo Research) and digested with CspI enzyme (ThermoFisher) according to manufacturer's protocol. Digestion products were then run and visualised on 2% agarose gel.

3.10.7 Optimisation of one-pot blocking/labelling reaction on model DNA

20 ng of 100 bp model DNA was glucosylated in 10 µL reaction containing 50 mM HEPES pH = 8.0, 10 mM MgCl₂, 1 mM DTT, 1 µL T4-βGT (ThermoFisher) and 150 µM UDP-N3-glucose alone or in combination with 1-100% of UDP-glucose at 37°C for 1 h. Subsequently DBCO-Cy5.5 was added to the reaction to final concentration of 0.3 mM) and samples were incubated for additional 2 hours at 37°C. Then reactions were purified on Zymo-Spin column (Zymo Research). Purified reaction products were run on 2% agarose gel. DNA was visualised using ChemiDoc Alexa Fluor 488 method while Cy5.5 labelling was visualised using Cy5.5 method.

3.10.8 Comparison of 5mC labelling reaction with NgTET1 and ccTET

5mC labelling reaction with NgTET1 and ccTET was tested on MC spike-in model DNA where 10% of all cytosines were methylated (Appendix 3, Supplementary Table 3-2). 40

ng of model DNA were labelled in blocked in 20 μ L reaction containing 50 mM HEPES pH = 8.0, 10 mM MgCl₂, 1 mM DTT, 2 mM L-ascorbic acid, 1 mM α -ketoglutarate, 75 μ M ammonium iron(II) sulfate, 0.35 μ M T4- β GT (ThermoFisher), 150 μ M UDP-N₃-glucose and 0.35 μ M, 0.75 μ M or 1.5 μ M NgTET1 or ccTET for 1 h at 37°C. Subsequently DBCO-Cy5.5 was added to the reaction to final concentration of 0.3 mM and samples were incubated for additional 2 hours at 37°C. Then reactions were purified on Zymo-Spin column (Zymo Research). Purified reaction products were run on 2% agarose gel. DNA was visualised using ChemiDoc Alexa Fluor 488 method while Cy5.5 labelling was visualised using Cy5.5 method.

3.10.9 Plasma Samples Collection and Processing

Healthy controls were recruited from John Radcliffe hospital (Ethical approvals IDs 16/YH/0247 and 18/WM/0237). Whole blood (5-10 mL) was collected into EDTA tubes and kept on ice. Plasma was separated within 1-4 hours after blood collection. Briefly, blood was transferred to Falcon™ 15 mL Conical Centrifuge Tubes and centrifuged at 1600 xg for 10 min at 4°C. Subsequently, top liquid layer was collected carefully without disturbing cellular pellet, transferred to DNA LoBind® Tubes (Eppendorf) and centrifuged at 16000 xg at 10 min at 4°C. Collected plasma was then divided into 1 mL aliquots and stored in DNA LoBind® Tubes (Eppendorf) at -80°C until cfDNA isolation. cfDNA was extracted from 1-4 mL plasma using Qiaamp Circulating Nuclei Acid Kit (Qiagen) and eluted in EB buffer. cfDNA concentration was quantified by Qubit Fluorometer (Life Technologies). Integrity and size distribution of cfDNA were assessed based on electrophoretic separation and visualisation on 2% E-Gel EX Agarose Gel (Invitrogen). All cfDNA samples included in the study exhibited distinct sharp band at ~

170 bp and did not show any signs of DNA degradation or contamination with genomic DNA. Isolated cfDNA was stored in DNA LoBind® Tubes (Eppendorf) at -20°C.

Pancreatic cancer, GBM, gastric cancer and colorectal cancer samples were historical cfDNA samples provided by Dr Chunxiao Song. Patients were recruited from West China Hospital (Ethical approval ID 2016-63). Blood and plasma collection, cfDNA isolation and quality control steps were performed by as described above for healthy controls.

3.10.10 5-methylcytosine pull-down in cfDNA

cfDNA samples were spiked with control oligos at concentration of 0.001 pg of each control oligo per 1 ng of cfDNA sample). 5hmC in the samples were blocked in 20 µL reaction mixture containing: 200 µM UDPG, 2 µL T4 β-glucosyltransferase (ThermoFisher), 50 mM HEPES pH = 8.0 and 2.5 mM MgCl₂ for 1h at 37°C. Then samples were purified on 1.8× AMPure XP beads according to manufacturer's protocol. For NgTet1 oxidation/ T4 β-GT labelling reaction samples were incubated in 50mM HEPES (pH 6.9), 75 µM Fe(NH₄)₂(SO₄)₂, 2mM ascorbic acid, 1mM α-KG, 10mM MgCl₂, 1mM dithiothreitol and 150 µM UDP-6-N3-Glc (Active Motif), 2 µL T4 β-glucosyltransferase (ThermoFisher) and 0.03-0.75 µM NgTET1 for 1 h at 37 °C. Then reaction was quenched by 30 min treatment with 1 µL proteinase K (NEB) at 37°C. For biotinylation, 2 mM DBCO-PEG4-biotin was directly added to the reactions, samples were incubated for additional 2 h at 37°C and purified on 1.8× AMPure XP beads. Labelled samples were subjected to End-repair and A-tailing reaction and Adapter ligation reactions with KapaHyper Plus kit according to manufacturer's protocol. Ligated libraries were labelled with 1 mM DBCO-PEG4-biotin in the presence of 50 mM HEPES buffer (pH 8) and 25 mM MgCl₂ for 2h at 37°C. Then samples were spiked-in with 10 µg salmon-sperm DNA and purified on Bio-Spin® Columns with Bio-Gel® P-30. Biotin-labelled DNA was

pulled-down on M270 Streptavidin beads (0.5 uL of beads per sample). Briefly, beads were first blocked with 10 µg of salmon-sperm DNA in 5 mM Tris pH = 7.5, 0.5 mM EDTA, 1 M NaCl and 0.2% Tween 20) for 30 min. Then, beads were incubated with the samples for 30 min and subsequently washed three times with each of following buffers:

- Buffer 1 - 5 mM Tris pH 7.5, 0.5 mM EDTA, 1 M NaCl and 0.2% Tween 20
- Buffer 2 - 5 mM Tris pH 7.5, 0.5 mM EDTA and 0.2% Tween 20
- Buffer 3 - 5 mM Tris pH 9.0, 0.5 mM EDTA, 1 M NaCl and 0.2% Tween 20
- Buffer 4 - 5 mM Tris pH 9.0, 0.5 mM EDTA and 0.2% Tween 20

Finally, beads were washed with nuclease-free water and captured DNA was amplified with KAPA HiFi HotStart ReadyMix for 14 cycles. PCR products were purified on AMPure XP beads and quantified with Qubit. Pair-end 75 bp sequencing was performed on the NextSeq instrument.

3.10.11 5mC pull-down validation by PCR

5mC pull-down efficiency and specificity were validated by end-point PCR or qPCR. For end-point PCR 1 ng of input library of enrichment library were PCR amplified by Taq DNA Polymerase (New England Biolabs) in 25 µL reaction with the corresponding primers according to manufacturer's protocol. Amplification products were separated and visualised on 2% agarose gel.

For qPCR, 1 ng of input and enrichment library were amplified by Fast SybrGreen MasterMix (ThermoFisher) according to manufacturer's protocol. To calculate fold enrichment, ΔC_t of particular spike-in control in input and enriched sample was calculated. Subsequently, fold enrichment value was calculated as ratio of ΔC_t of methylated and ΔC_t of unmodified spike-in controls (fold enrichment over C) or ΔC_t of

methyated and ΔCt of hydroxymethylated spike-in controls (fold enrichment over 5hmC).

Sequences of primers and final products and annealing temperatures for primers are listed in Appendix 3 Supplementary Table 3-7.

3.10.12 Data processing

Bowtie2 v2.2.6-2 was used to align FASTQ sequences to UCSC/hg19 genome (Langmead and Salzberg, 2012). Aligned sequences were then filtered with samtools-0.1.19 (view -f 2 -F 1548 -q 30 and rmdup) to remove duplicated or poor-quality reads (Li *et al.*, 2009). BigWig files were produced with Deeptools (Ramírez *et al.*, 2014). Numbers of fragments in genomic bins, promoters, genebodies, CGIs and enhancers were calculated by bedtools (Quinlan and Hall, 2010). Batch effect in the data was removed by sva package (Leek *et al.*, 2014). For analyses reads mapped to chromosome X and Y were excluded. Efficiency and specificity of pull-down were validated based on control sequences alignment.

3.10.13 Differential methylation analysis

5mC Log2 Fragments per kilo base per million mapped reads (FPKM) were calculated for genomic regions in R. PCA, correlation analysis, hierarchical clustering, GO analysis and heatmaps were prepared in R. Differential methylation analysis in promoters and genebody regions for pancreatic cancer was performed by *Limma* package (Ritchie *et al.*, 2015). Differentially methylated genes were defined based on p value lower than 0.05. For promoters the nearest gene was identified by bedtools (Quinlan and Hall, 2010). Normalised 5mC coverage for metagene plots was calculated by bamCoverage and computeMatrix functions in deeptools (Ramírez *et al.*, 2014) and plotted in R.

*Progress is made by trial and failure;
the failures are generally a hundred times more numerous than the successes;
yet they are usually left unchronicled.*

William Ramsay

4 Development of Tet-assisted Pyridine Borane Sequencing (TAPS) for base-resolution DNA methylation analysis

4.1 Introduction

Over the last decades, a wide range of new technologies has been developed to enable analysis of cytosine modifications in DNA (Plongthongkum, Diep and Zhang, 2014). In general, next generation sequencing approaches which profile 5mC and 5hmC genome-wide are the most accurate and comprehensive way to study methylome (Plongthongkum, Diep and Zhang, 2014). As reviewed in the introduction, current methylation sequencing technologies have their own limitations and there is still need for development of new methods for quantitative, sensitive and cost-effective sequencing of epigenetic marks in DNA.

In enrichment-based technologies, methylated regions or CpG-rich regions are captured by restriction digestion (Meissner *et al.*, 2005; Wang *et al.*, 2013) or affinity enrichment (Song *et al.*, 2011; Taiwo *et al.*, 2012; Zhang *et al.*, 2013). These strategies significantly reduce sequencing costs when compared to whole-genome sequencing but simultaneously they have limited resolution and quantitative potential (Ziller *et al.*, 2013; Plongthongkum, Diep and Zhang, 2014; Breiling and Lyko, 2015).

In contrast, whole-genome methylation sequencing enables 5mC mapping at the base resolution level and is a fully quantitative and high resolution way to study methylome in a site-specific manner (Susan *et al.*, 1994; Plongthongkum, Diep and Zhang, 2014; Y.

Liu *et al.*, 2019). Although this approach is more expensive, it overcomes limitations of enrichment-based techniques in the areas where very detailed information about methylation status is desirable. Recent findings have suggested that methylation status of single CpG has a functional and biological effect, which emphasises great significance of base-resolution methylome analysis for complete understanding of methylation role in genome regulation (Kim *et al.*, 2003; Zhang and Pradhan, 2014; Xu *et al.*, 2019).

For a long time, the only way and the gold standard for base-resolution methylome analysis was bisulfite sequencing (BS-seq) (Frommer *et al.*, 1992; Susan *et al.*, 1994) and other methods based on bisulfite treatment such as Oxidative Bisulfite Sequencing (oxBS-seq) (Booth *et al.*, 2013; Peng, Xia and Yi, 2016) or Tet-Assisted Bisulfite Sequencing (TAB-seq) (Yu *et al.*, 2012). BS-seq relies on chemical conversion of unmodified cytosine to uracil, which is then read as thymine after PCR amplification. In turn, both 5mC and 5hmC remain intact after this treatment and are read as cytosine during sequencing.

Despite being widely used for methylome studies, bisulfite method has serious disadvantages. First, bisulfite treatment is a harsh chemical reaction and degrades up to 95% of input DNA (Tanaka and Okamoto, 2007). Moreover, as most cytosines in the genome are unmodified and are converted to thymine after bisulfite treatment, final sequence contains as low as 1% cytosines. This leads to severe drop in sequence complexity and causes subsequent problems during Illumina sequencing (Olova *et al.*, 2018). As final result, BS-seq has lower sequencing quality and decreased mapping rate compared to standard whole-genome sequencing. Therefore, further technological improvements in a field are still required.

This chapter describes development of new, base-resolution method for methylome sequencing later called Tet-assisted Pyridine Borane Sequencing (TAPS) and its application for whole-genome methylome sequencing in mouse embryonic stem cells.

4.2 Motivation

We aimed to fill a technological gap existing in epigenetic technologies and develop a new, base-resolution methylome sequencing method as an alternative to bisulfite sequencing. Importantly, we wanted to address major problems and limitations of the current gold standard.

Our first goal was to design a milder chemical reaction to reduce DNA damage and facilitate genome-wide methylation sequencing in challenging, low-input applications such as cfDNA or single-cell analysis. Second, we wanted to develop a direct way for 5mC and 5hmC detection to avoid technical issues during Illumina sequencing caused by low sequence complexity characteristic for bisulfite-converted libraries (Olova *et al.*, 2018).

To achieve these goals, we had to develop a new chemistry able to induce base change directly at modified sites and further combine it with next generation sequencing protocol. Finally, we applied newly developed TAPS protocol to methylation sequencing in mouse embryonic stem cells (mESC) and compared it with the current gold standard - bisulfite sequencing.

4.3 Borane compounds reduce 5caC to DHU

5-Methylcytosine (5mC) is chemically stable and lowly-reactive nucleotide and a chemical reaction directly targeting this nucleotide and inducing base change has yet to be discovered (Hofer, Liu and Balasubramanian, 2019; Zeng, He and Yi, 2019). In

contrast, a new chemistry targeting its more reactive oxidation product - 5-formylcytosine (5fC) and inducing C-to-T transition of modified sites has been recently published and applied to base-resolution sequencing of 5fC in the genome (Xia *et al.*, 2015; Zhu *et al.*, 2017). This method was further expanded to 5hmC mapping as selective chemical oxidation of 5hmC to 5fC had been developed (Zeng *et al.*, 2018; Zeng, He and Yi, 2019). However, to date there is no selective and efficient method to oxidise 5mC to 5fC. In turn, both 5mC and 5hmC can be efficiently *in vitro* oxidised by TET enzymes to their final oxidation product 5caC (Yu *et al.*, 2012; Pais *et al.*, 2015; Liu *et al.*, 2017). Since 5caC is more reactive than 5mC (Hofer, Liu and Balasubramanian, 2019), we envisioned that similar to 5fC, 5caC could be chemically converted to introduce 5caC-to-T base change.

To identify chemicals that could react with 5caC, we designed 11mer DNA oligonucleotide containing a single 5caC and screened it against a panel of borane-containing compounds (Table 4-1).

Products of 5caC reduction by the borane compounds were characterised by matrix-assisted laser desorption/ionization mass spectroscopy (MALDI). This initial screening revealed that some borane-containing compounds could react with 5caC and induced decrease in molecular weight of the model DNA by 41 Da (Figure 4-1). We confirmed at a single nucleotide level, that the shift in molecular weight observed after the reactions with pic-borane and pyridine borane was the result of a previously unknown reaction of reductive decarboxylation/deamination of 5caC to dihydrouracil (DHU) and that 5fC could be also converted to DHU in a similar reductive deformylation/deamination reaction. We selected two the most promising compounds: 2-picoline borane (pic-borane, pic-BH₃) and pyridine borane (PyBr) for further optimisation.

Table 4-1. Summary of MALDI screening of borane-containing compounds panel. Experiment was performed by Dr Yibin Liu. 2-picoline borane, borane pyridine, tert-butylamine borane, and ammonia borane could completely convert 5caC to DHU in 11mer model double-stranded DNA (dsDNA) but only pyridine borane and 2-picoline borane were active on longer, model dsDNA. Sodium borohydride, ethylenediamine borane, dimethylamine borane and sodium triacetoxyborohydride yielded incomplete conversion between 20-60%. No detectable products were measured with morpholine borane, 4-methylmorpholine borane, trimethylamine borane, and cyclohexylamine borane.

Compound	5caC conversion in 11mer dsDNA	Activity on long dsDNA
Pyridine borane	100%	Yes
2-picoline borane	100%	Yes
Tert-butylamine borane	100%	No
Ammonia borane	100%	No
Sodium borohydride	60%	Not tested
Ethylenediamine borane	30%	Not tested
Dimethylamine borane	30%	Not tested
Sodium triacetoxyborohydride	20%	Not tested
Morpholine borane	Product not detected	Not tested
4-methylmorpholine borane	Product not detected	Not tested
Trimethylamine borane	Product not detected	Not tested
Dicyclohexylamine borane	Product not detected	Not tested

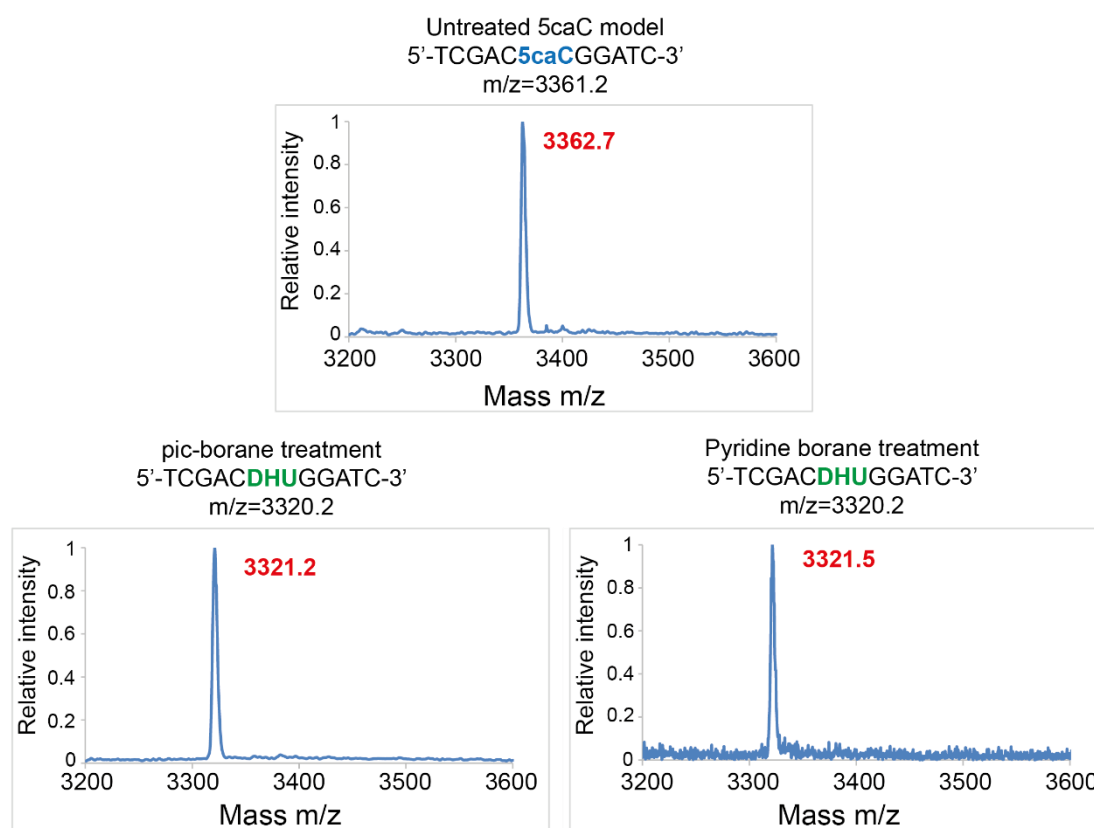


Figure 4-1. MALDI characterization of 5caC-containing 11mer model DNA treated with pic-borane and pyridine borane. After reaction shift in mass [m/z] corresponding to 5caC reduction to DHU was observed. Calculated mass was shown in black, and observed mass was shown in red. Experiment was performed by Dr Yibin Liu.

We envisioned that Ten-eleven translocation (TET) enzyme oxidation could be combined with newly discovered chemistry to develop novel methylome sequencing method. We named it TET-assisted Pyridine/pic-Borane Sequencing (TAPS) (Figure 4-2).

As a proof-of-concept experiment, we repeated borane reduction on double-stranded 222 bp CpG-methylated model DNA, previously oxidised by NgTET1 enzyme and confirmed in Sanger sequencing that our treatment indeed could introduce C-to-T transition in modified sites as DHU was read as thymine after PCR amplification (Figure 4-3). This finding enabled us to start working on further optimisation of TAPS method for whole genome methylome sequencing applications.

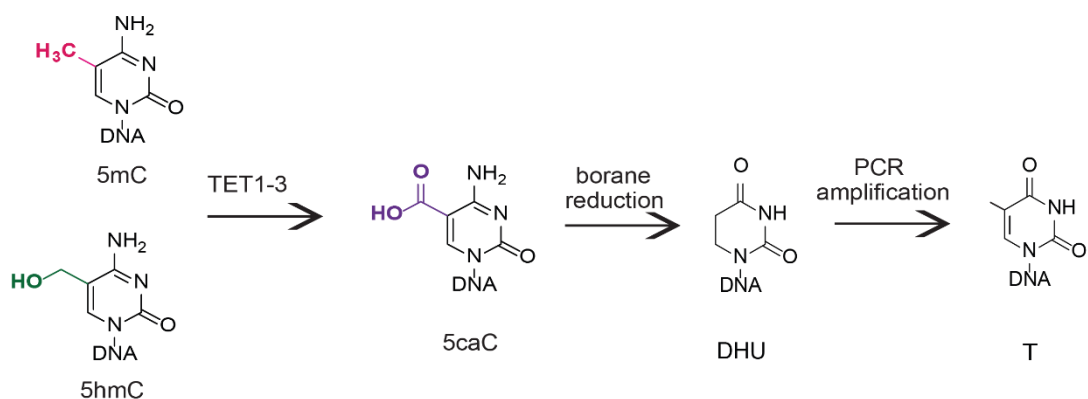


Figure 4-2. Overview of TAPS method design. 5mC and 5hmC in DNA are oxidized to 5caC by TET enzyme and then reduced to DHU with borane-containing compounds. DHU induces base change during PCR amplification and is read as T, which enables direct detection of modifications.

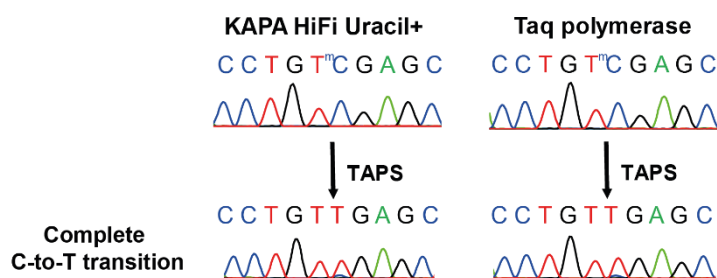


Figure 4-3. Sanger sequencing results confirming C-to-T after PCR amplification of 222 bp model DNA treated with TAPS. After the treatment 5mC was converted to DHU and read as T by KAPA HiFi Uracil plus polymerase and Taq polymerase. Experiment was performed once by Dr Yibin Liu.

4.4 Recombinant TET enzymes for TAPS

Efficient and complete *in vitro* oxidation of 5mC and 5hmC to 5caC is an important step in the TAPS protocol. Therefore, the next step in our method development pathway was to produce a highly active and pure recombinant TET enzyme. We had already produced active full length TET1 enzyme from *Naegleria guberi* (NgTET1) for cfMeSeal method (Expression and purification details - section 3.4, activity validation - 3.8). In addition, we aimed to produce recombinant catalytic domain of mouse Tet1 (mTet1CD) previously used in TAB-seq (Yu *et al.*, 2012).

mTet1CD enzyme was initially expressed in adherent HEK293T cells by transient transfection but this protocol was further simplified by optimising protein expression in suspension Expi293F cells. mTet1CD was purified in a single-step Flag-tag purification protocol at high purity and quality (Figure 4-4). However, overall protein yield was lower than for NgTET1 and for most purification batches did not exceed 0.5 mg of protein from 1 l cells.

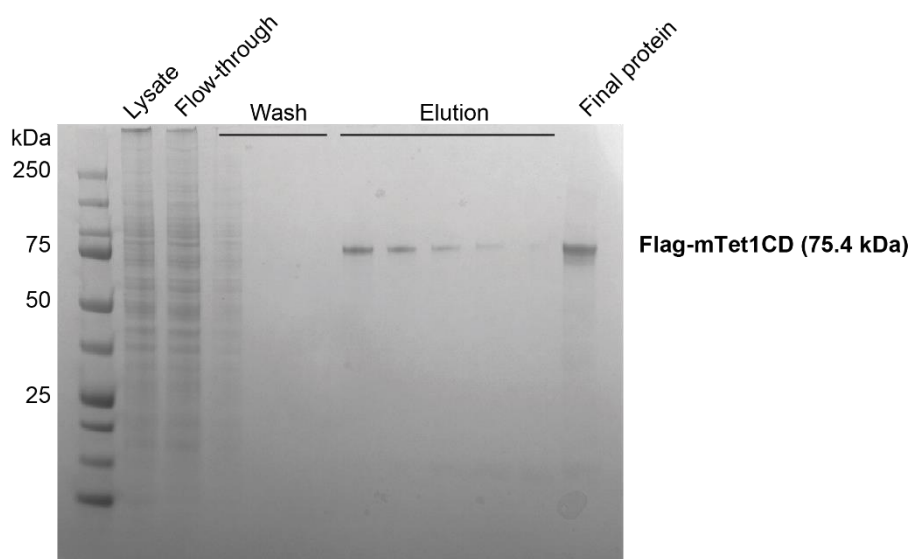


Figure 4-4. Representative SDS-PAGE electrophoresis gel of mTet1CD purification. Lysate was protein sample before passing through ANTI-FLAG M2 gel. Flow-through contained unbound fractions after passing through the resins. Wash fractions contained contaminant proteins removed after washing with TBS. Elution fractions contained eluted proteins after washing with buffer containing Flag-peptide. All elution fractions were collected and concentrated. Final protein was concentrated mTet1CD after collection all elution fractions.

We assessed enzymatic activity of recombinant mTet1CD and compared it with NgTET1 proteins by oxidation of 11mer model DNA oligonucleotide containing single, hemimethylated CpG site. On this model, both enzymes completely converted 5mC to 5caC (Figure 4-5). Additional peaks on MALDI spectrum after DNA oxidation by NgTET1 correspond to the products of thymine oxidation – formyluracil and carboxyluracil that are the products of off-target oxidation of thymine by NgTET1 previously described in the literature (Pais *et al.*, 2015).

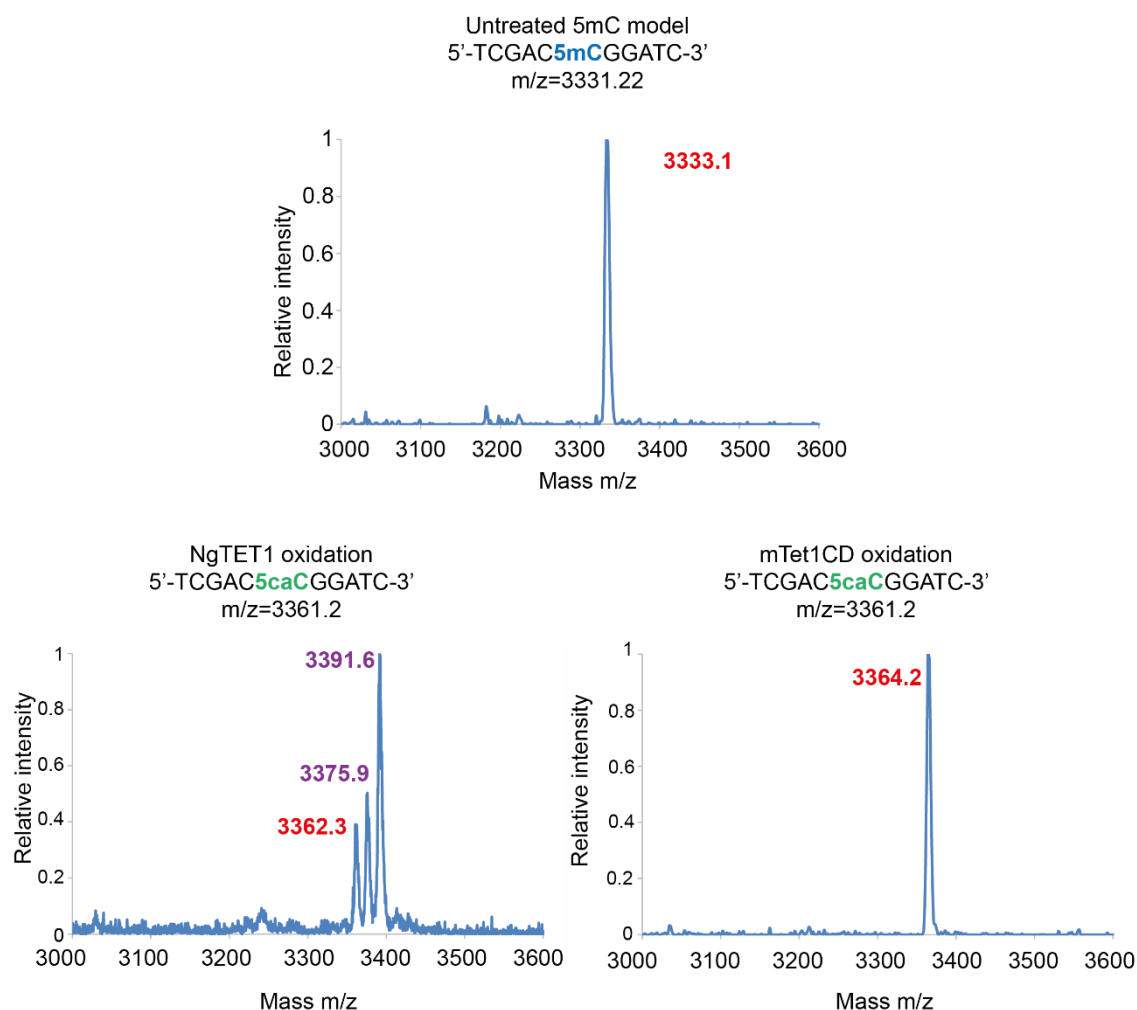


Figure 4-5. Representative results of TET activity assay measured by MALDI. MALDI characterization of 5 μ M 11mer 5mC DNA model before and after oxidation with 5 μ M NgTET1 or 4 μ M mTet1CD. Calculated mass of 11mer after complete 5mC to 5caC oxidation was shown in black, and observed mass was shown in red. Additional peaks after oxidation with NgTET1 shown in purple are results of off-target T oxidation to fdU (3375.9) and cadU (3391.6).

4.5 Spike-in controls for 5mC conversion monitoring

We next aimed to optimise TAPS protocol on genomic DNA to enable bisulfite-free, base-resolution sequencing of 5mC and 5hmC. To assess 5mC conversion and false positive rate in genomic samples we designed a set of DNA controls of known sequence (called here spike-in controls), which had modified or unmodified cytosines in the specific positions. The spike-in controls could be added to the sample in a small proportion and base change at the known modified or unmodified positions could be monitored to assess performance of the method.

All spike-in controls that were used during TAPS development are listed below:

- CpG-methylated lambda DNA to measure 5mC conversion (true positive rate) in CpG context
- Unmethylated 2kb amplicon to measure unspecific conversion (false positive rate) of unmodified cytosines
- Synthetic 160mer with single methylated and hydroxymethylated cytosines at various sequence contexts (N5mCNN and N5hmCNN)

4.6 5mC conversion can be monitored by restriction digestion assay

For efficient optimisation of TAPS on genomic samples, we needed a robust, simple and cheap assay to monitor 5mC conversion during optimisation process. We designed an assay based on restriction digestion of a control amplicon by Taq α I enzyme recognizing TCGA sequence. In this assay, an amplicon generated on unconverted DNA template can be cleaved by the enzyme while a PCR product amplified on TAPS-treated DNA cannot be digested due to 5mC-to-T transversion and subsequent loss of Taq α I restriction site. We initially used 222 bp sequence in lambda spike-in containing 5 CpG sites including one inside TCGA Taq α I digestion site. This fragment could be PCR amplified from DNA sample after conversion and further used for Taq α I restriction digestion assay. Digestion products were then run on 2% agarose gel and TAPS conversion level was quantified based on the intensity of digested and undigested bands (Figure 4-6).

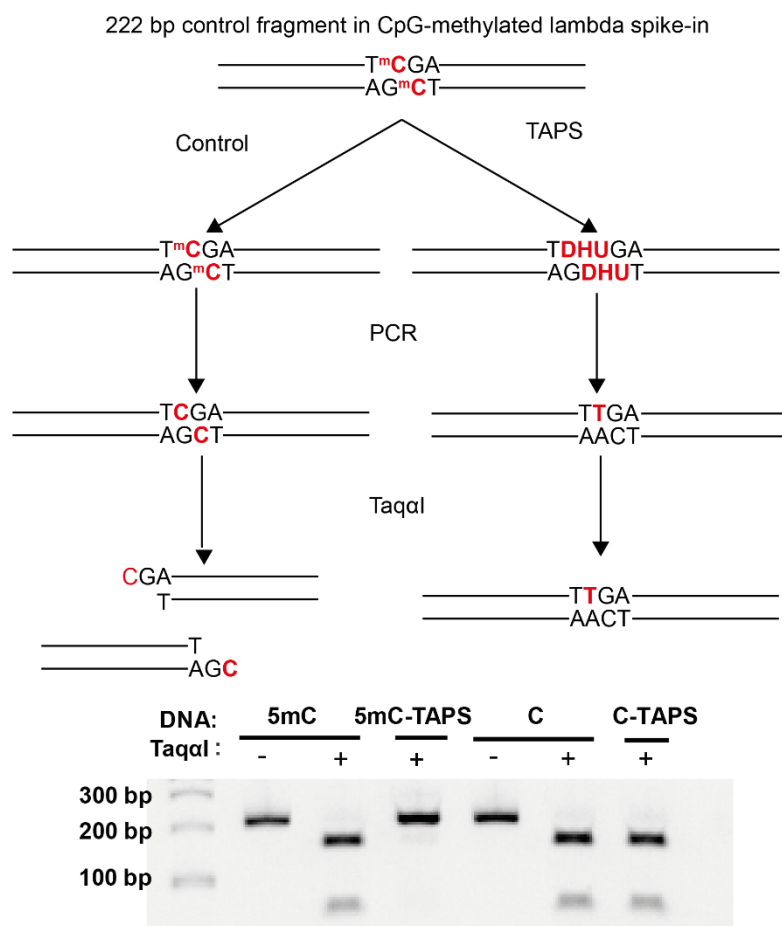


Figure 4-6. Overview of TaqI assay to test TAPS-induced base change in CpG methylated 222 bp model DNA. A PCR-amplified 222 bp model DNA with TaqI restriction site in the middle can be cleaved, whereas the amplified product of 5mC-TAPS stayed intact, suggesting loss of the restriction site due to C-to-T transition. TAPS did not result in C-to-T transition on the unmethylated cytosine since C-TAPS was cleaved in the same way as the original untreated C. Experiment was performed once by Dr Yibin Liu.

4.7 Pic-borane reduces oxidised 5caC in a genome

After the initial promising results of TAPS on the model DNA, we proceeded with method optimisation on genomic DNA and development of whole-genome sequencing TAPS protocol.

We found NgTET1 enzyme easier to produce at high yield with very similar activity to mTET1CD and pic-borane as easier to handle due to its lower toxicity. Therefore, we focused on the combination of NgTET1 oxidation and pic-borane reduction at the initial stages of the method development.

We applied pic-borane reduction conditions that were previously shown to work on model DNA to mESC genomic DNA (Figure 4-7, a) and confirmed efficient oxidation and reduction of modified cytosines by HPLC-MS/MS and Taq α I assay (Figure 4-7, b). We thus used these conditions to prepare the next generation sequencing library (Figure 4-7, c) and sequenced it on Illumina platform.

Surprisingly, 5mC conversion rate estimated from sequencing data was significantly lower than values measured by HPLC/MS-MS since we achieved 70.51% of 5mC conversion in CpG context based CpG-methylated lambda (Figure 4-7, d). We found that 5mC conversion in methylated lambda ranged from few to almost 100% depending on GC content of sequences surrounding particular CpG site, with the lowest conversion in GC-rich regions (Figure 4-7, e).

These results suggested that under the tested conditions DNA fragments with high GC content and potentially higher local melting temperature were not efficiently converted. We hypothesised that low conversion was caused rather by incomplete pic-borane reduction than by an inefficient oxidation of 5mC and 5hmC by NgTET1 and therefore we focused on further optimisation of pic-borane reaction conditions.

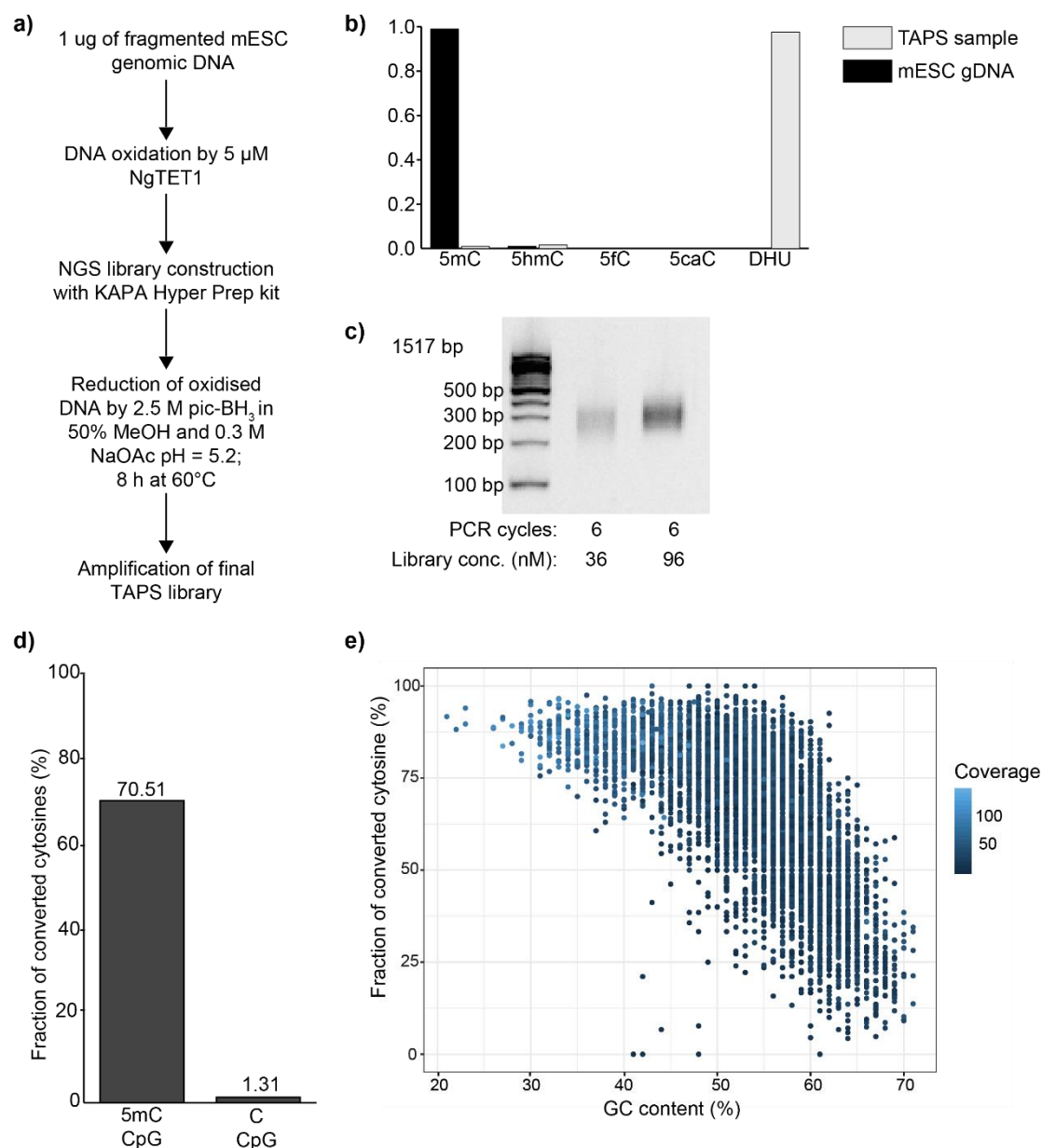


Figure 4-7. Preparation of first TAPS sequencing library. (a) General overview of TAPS library preparation. (b) HPLC-MS/MS quantification of relative modification levels in the mESC genomic DNA sequencing sample before and after NgTET1 oxidation and pic-borane reduction. HPLC-MS/MS was performed by Fang Yuan. (c) Agarose gel of final TAPS libraries constructed with KAPA HyperPrep double strand library kit and Methyl-seq single strand library kit after post-amplification clean-up (d) Conversion rate in TAPS library based on spike-in controls. Conversion of 5mC was calculated based on CpG-methylated lambda. False positive rate (conversion of unmodified cytosine) was quantified based on unmodified 2kb amplicon. Sites covered by less than 3 sequencing reads were excluded from analysis. (e) Negative correlation between 5mC conversion rate of single, methylated CpG and GC content in 200 bp window around this site (Pearson correlation: -0.68). Sites covered by less than 3 sequencing reads were excluded from analysis.

Because of discrepancies in conversion rates between regions with low and high GC content, we designed two additional testing probes for amplification on methylated lambda control spike-in for Taq α I conversion assay: 216 bp low GC model (38% GC content) and 194 bp high GC model (60.8% GC content).

We assumed that higher reaction temperature and addition of DMSO could both help to relax GC-rich fragments of DNA and facilitate their conversion. We then replaced 50% MeOH used in previous reduction conditions with DMSO and tested pic-borane reaction at different temperatures, pH and reaction times. Results of Taq α I assay for this condition screening confirmed that all tested conditions efficiently converted 5mC in low GC DNA fragments while they exhibited significant differences in conversion efficiency of 5mC in DNA fragments with high GC content. In summary, 5mC conversion in these regions was significantly improved when pic-borane reduction was performed at the higher temperature and in the presence of DMSO (Figure 4-8).

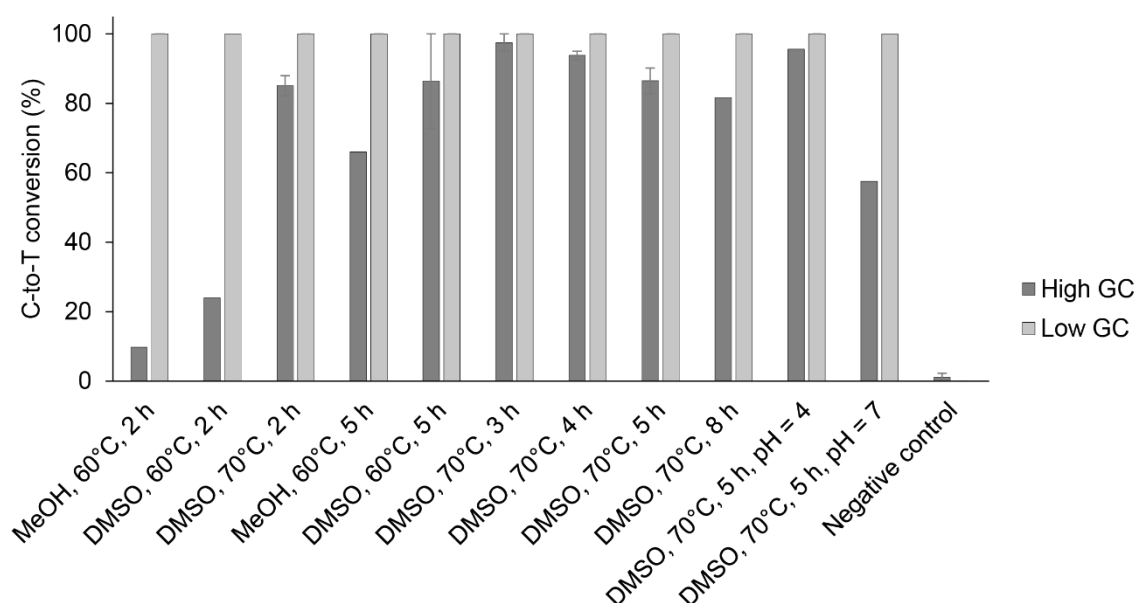


Figure 4-8. The summary of pic-borane reaction optimisation conditions tested to improve the C-to-T conversion in high GC regions. The C-to-T conversion was estimated by Taq α I-digestion assay of low and high GC density sequences amplified from CpG methylated lambda spike-in control after TET oxidation and pic-borane conversion. If not specified, reactions were performed

at pH = 5.2. Error bars representing +/- standard error are shown for condition for which two or three independent technical replicates were prepared.

We selected pic-borane reaction performed in the presence of 50% DMSO for 3 h at 70°C as the most promising condition to prepare Illumina sequencing library (Figure 4-9, a-b). High 5mC conversion in low and high GC spike-in controls in this sample was confirmed by TaqαI assay prior sequencing (Figure 4-9, c). Illumina sequencing confirmed improved conversion of 5mC at CpG context of 87.3% based on methylated lambda DNA. Simultaneously, we observed an undesired increase in unspecific conversion of unmodified cytosine that resulted in a false-positive rate to be as high as 1.83% (Figure 4-9, d). Additionally, we investigated a relationship between GC content and 5mC conversion in lambda DNA and confirmed that the optimised reaction improved 5mC conversion in GC-rich regions and that under these conditions conversion rate did not negatively correlate with GC content of surrounding DNA (Pearson correlation: -0.08; Figure 4-9, e)

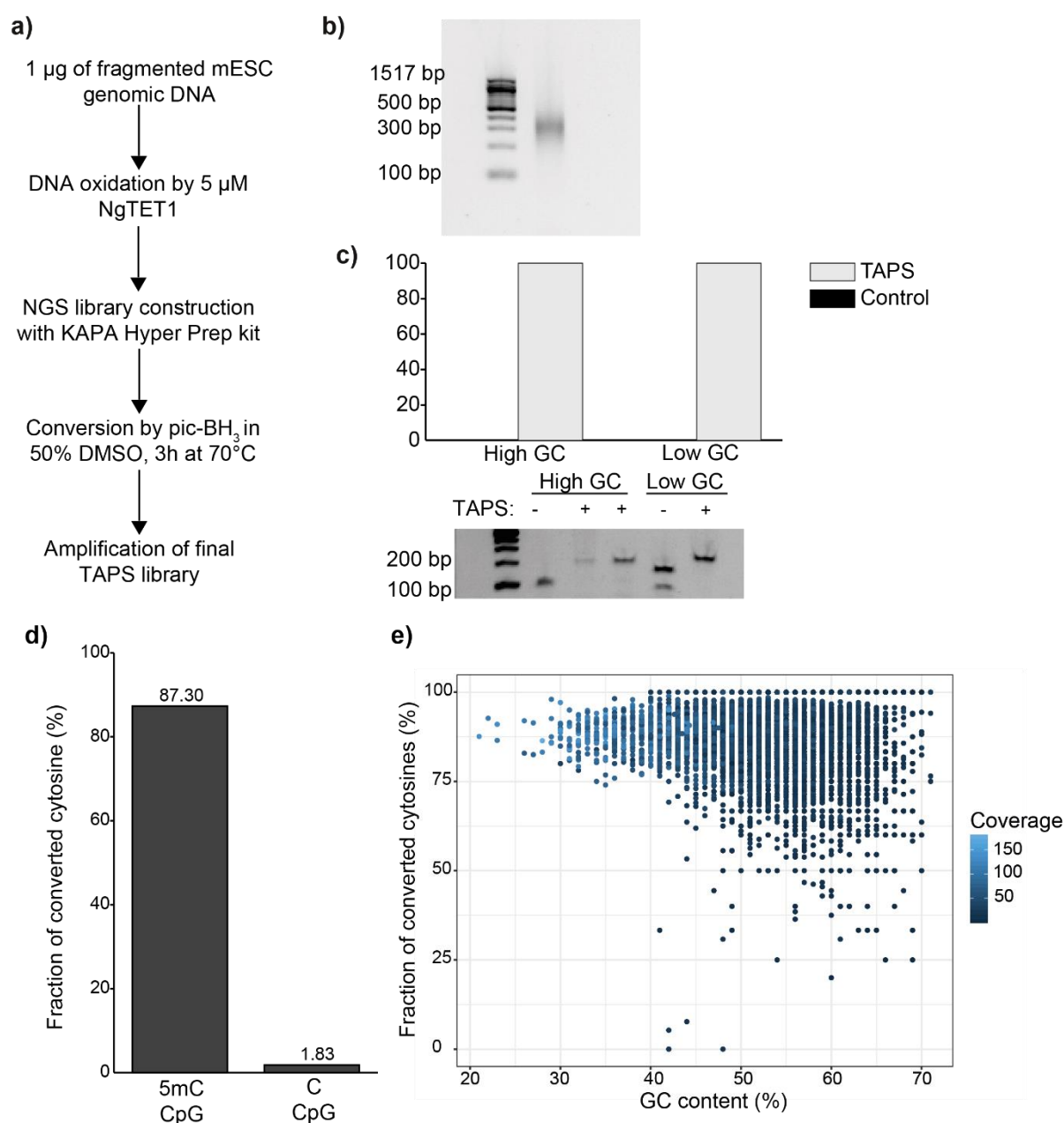


Figure 4-9. Library preparation for optimised pic-borane reaction conditions. (a) General overview of TAPS library preparation according to the protocol optimised after second optimisation round. (b) Agarose gel of final TAPS library constructed with KAPA HyperPrep double strand library kit after post-amplification clean-up. (c) Results of TaqI-digestion assay to validate 5mC conversion in TAPS library prior to sequencing. Restriction digestion confirmed complete C-to-T transition caused by TAPS both on high and low GC validation probe. (d) 5mC conversion rate in TAPS library based on spike-in controls. Conversion of 5mC was calculated based on CpG-methylated lambda. False positive rate (conversion of unmodified cytosine) was quantified based on unmodified 2kb-3 amplicon. Sites covered by less than 3 sequencing reads were excluded from analysis. (e) Lack of correlation between 5mC conversion rate of single, methylated CpG in lambda DNA and GC content in 200 bp window around this site (Pearson correlation: -0.08). Sites covered by less than 3 sequencing reads were excluded from analysis.

In addition, we demonstrated that TAPS could be applied to low input DNA samples and produced sequencing libraries at good yield and at the expected size range (Fig. 4-10, a-c).

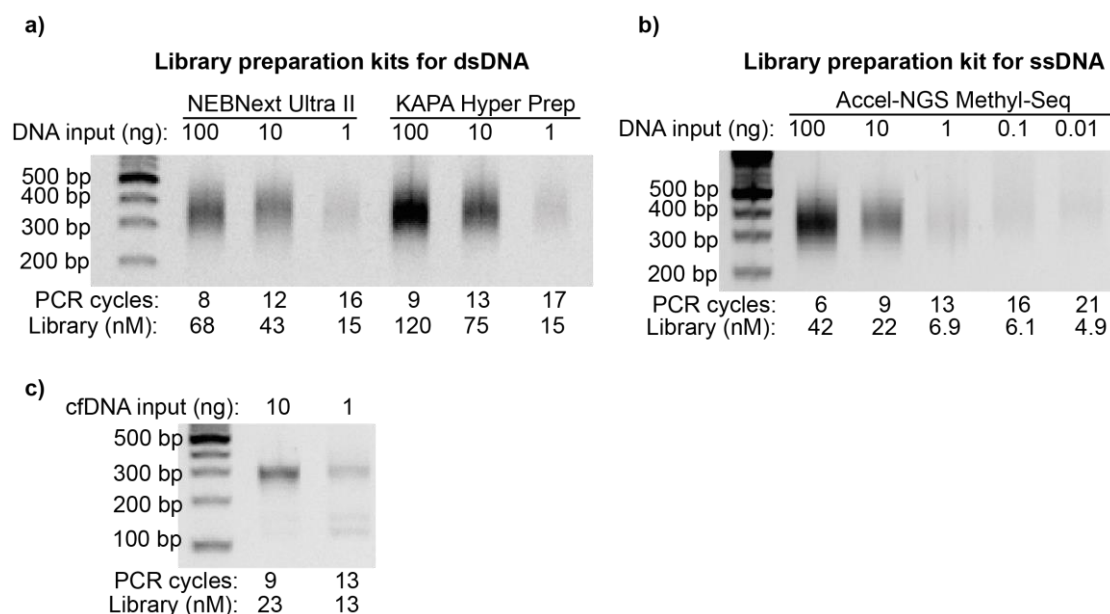


Figure 4-10. Low-input gDNA and cell-free DNA TAPS libraries prepared with dsDNA and ssDNA library preparation kits according to protocol established after second optimisation round. (a) Sequencing libraries were successfully constructed with down to 1 ng of mESC gDNA with dsDNA library kits NEBNext Ultra II or KAPA Hyperplus kits. (b) Single-strand TAPS libraries were prepared with Accel-NGS Methyl-Seq kit from as low as 0.01 ng of mESC gDNA. (c) CfDNA TAPS libraries were prepared with Accel-NGS Methyl-Seq kit with down to 1 ng of cfDNA.

Taken together, we successfully combined NgTET1 oxidation and pic-borane reduction with Illumina library preparation to develop TAPS and generated genomic and cfDNA sequencing libraries at the high yield and expected molecular weight and with good sensitivity for 5mC detection. We applied developed TAPS protocol to the genomic samples and demonstrated a good potential of this method for low input application. However, due to high false positive rate (1.83%) and relatively high false negative rate (12.7%) further optimisation was required to ensure highly sensitive and accurate detection of modified cytosines in the genome.

4.8 mTet1 oxidation followed by pyridine borane reduction improves 5mC detection in a genome

5mC conversion below 90% and high false positive rate observed for NgTET1 and pic-borane TAPS conditions could be attributed both to the enzyme performance and chemical reduction. Therefore, we extended our optimisation to the second hit of initial compound screening - pyridine borane (PyBr) as well as to mTet1CD as an oxidizing enzyme. We first discovered that pyridine borane conditions significantly improved DNA recovery after post-conversion clean up and enabled at least 5-fold decrease of input DNA. (Figure 4-11).

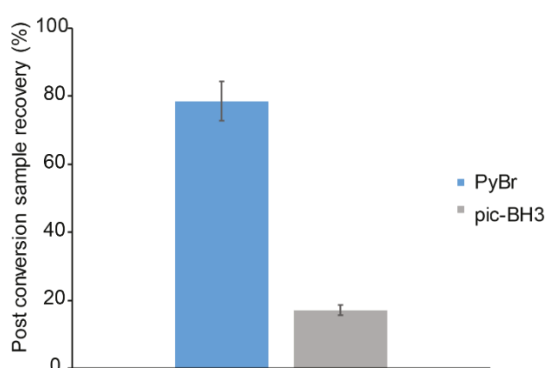


Figure 4-11. Comparison of post-conversion sample recovery after treatment with pic-borane and pyridine borane. Data shown as mean $\pm$ SE of three replicates.

For more accurate assessment of 5mC conversion and monitoring of false positive rate, we used shallow Illumina sequencing to validate the most promising conditions identified by Taq α I assay. We performed first screening to test initial conditions of pyridine borane reduction developed by Dr Yibin Liu (2 M pyridine borane at 90 mM NaOAc pH = 5.2 at 50°C) at the different reaction times (16 h, 12 h and 8 h) in combination with oxidation by mTet1CD or NgTET1. We found that mTet1CD-based condition performed better than NgTET1 and that longer pyridine borane reaction time had a little effect on conversion. Although, none of the tested conditions improved 5mC conversion to the

level higher than previously observed for pic-borane reduction, all of them exhibited reduced false-positive rate (Figure 4-12).

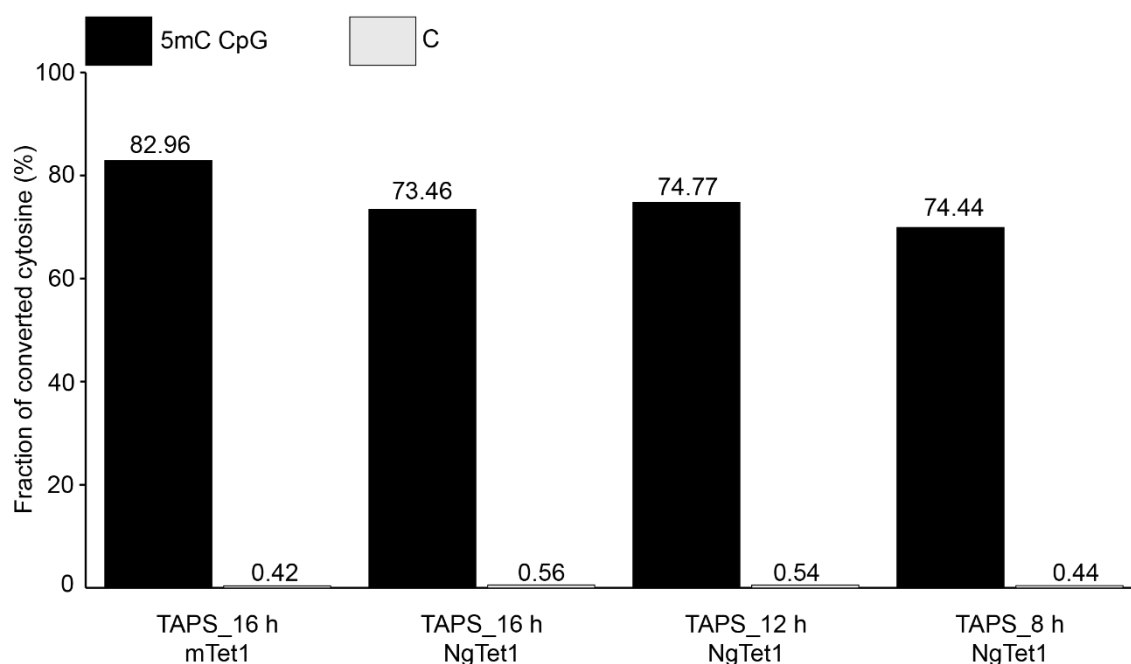


Figure 4-12. Results of test sequencing run to screen different TET oxidation and pyridine borane reduction conditions. 5mC conversion rate in CpG context was estimated based on CpG-methylated lambda DNA. False positive rate of TAPS was calculated from unmodified 2kb spike-in. Sites covered by less than 3 sequencing reads were excluded from analysis.

Due to improved DNA recovery and reduced false positive rate, we aimed to optimise pyridine borane reaction conditions further to increase conversion rate. Potential mechanism of 5caC to DHU conversion during pyridine borane reaction is reductive decarboxylation/deamination. The deamination reaction is known to be facilitated by acidic pH (Shapiro, DiFate and Welcher, 1974). Therefore, we predicted that ensuring low pH during conversion could help to increase 5caC reduction efficiency. As pyridine borane is a basic compound, this could be achieved by reduction of compound concentration, decrease of reaction buffer pH and/or increase of buffer concentration. We performed condition screening with TaqαI assay and confirmed that both decreased pH and increased concentration of sodium acetate buffer improved 5mC conversion with pH = 4.3 producing the best results (Figure 4-13, a). We then demonstrated that reaction was

the most efficient when performed at 37-40°C and that 1 M pyridine borane was the most optimal amount of compound (Figure 4-13, b). Finally, we tested these TAPS conditions in combination with library preparation protocol and found that reaction performed at 1M PyBr and 600 mM sodium acetate pH = 4.3 resulted in the highest 5mC conversion rate (Figure 4-13, c)

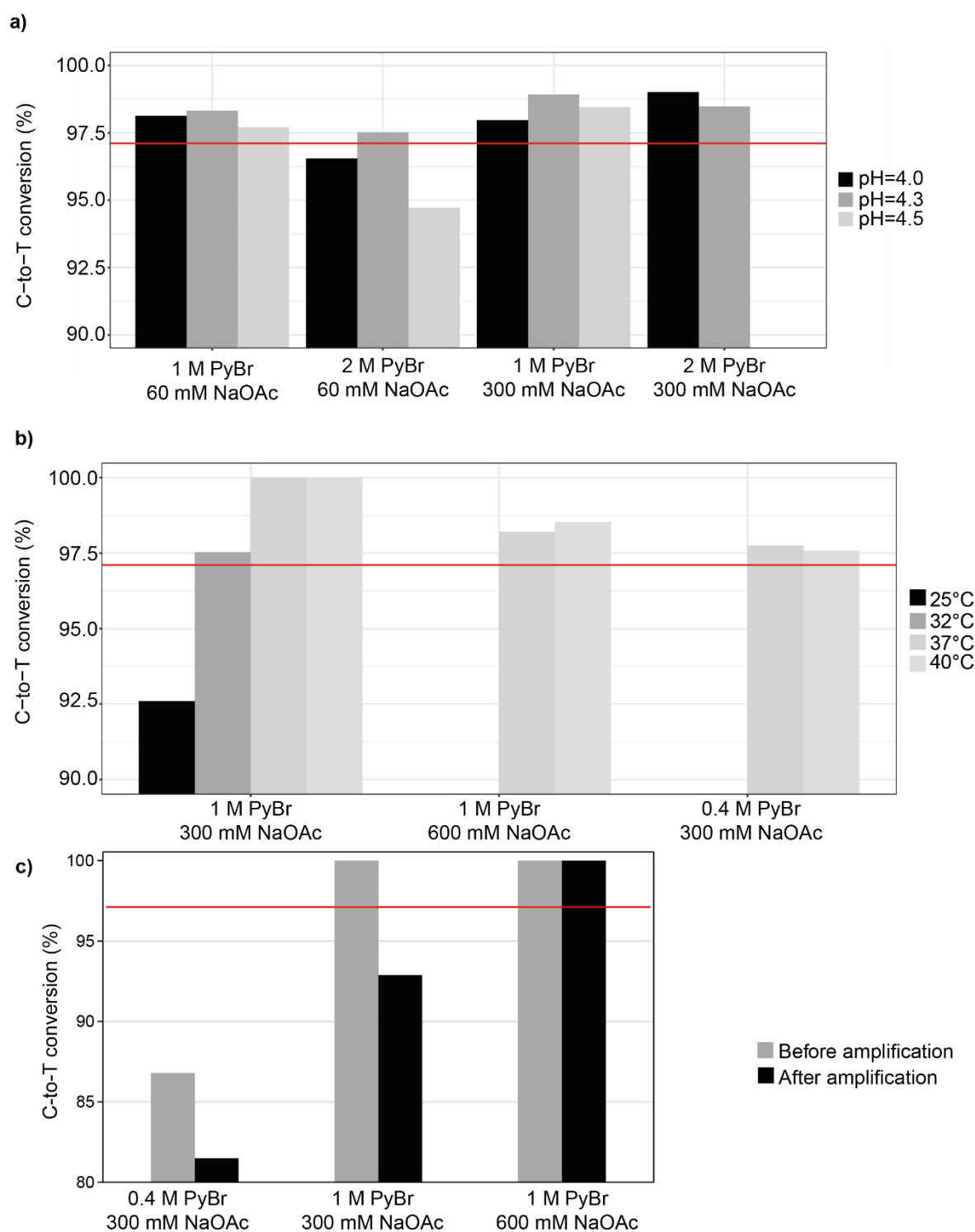


Figure 4-13. Optimisation of TAPS protocol with pyridine borane compound. (a) The summary of pyridine borane reaction optimisation on model 194 bp DNA. Taq α I digestion assay results of optimisation of PyBr reaction conditions on model 194mer DNA. The C-to-T conversion was estimated by Taq α I-digestion assay. Red horizontal line represents 5mC conversion in the sample treated according to the pic-borane protocol that was used as a reference point to assess tested conditions. (b) The summary of pyridine borane reaction follow-ups at pH = 4.3 tested in mESC genomic DNA. The C-to-T conversion was estimated by Taq α I-digestion assay. Red horizontal line represents 5mC conversion in the sample treated according to the pic-borane protocol that was used as a reference point to assess the conditions tested. (c) Taq α I digestion assay results of the best pyridine borane conditions tested with NGS library preparation protocol before and after library amplification. Red horizontal line represents 5mC conversion in the sample treated according to the pic-borane protocol that was used as a reference point to assess the conditions tested.

We combined these reduction conditions in combination with NgTET1 and mTet1CD oxidation, applied to the genomic samples and validated by shallow sequencing (Figure 4-14, a). Sequencing revealed that the combination of mTet1CD with pyridine borane reduction performed better and resulted in the highest conversion of 5mC (96.5% at CpG sites) and very low false positive rate of 0.23% (Figure 4-14, b). Similar to the previous optimised pic-borane reduction conditions, pyridine borane conversion was not affected by GC content (Figure 4-14, c). Owing to the high conversion rate and low positive rate mTet1CD oxidation followed by pyridine borane conversion in the presence of 600 mM NaOAc (pH 4.3) were selected as the final condition for TAPS protocol.

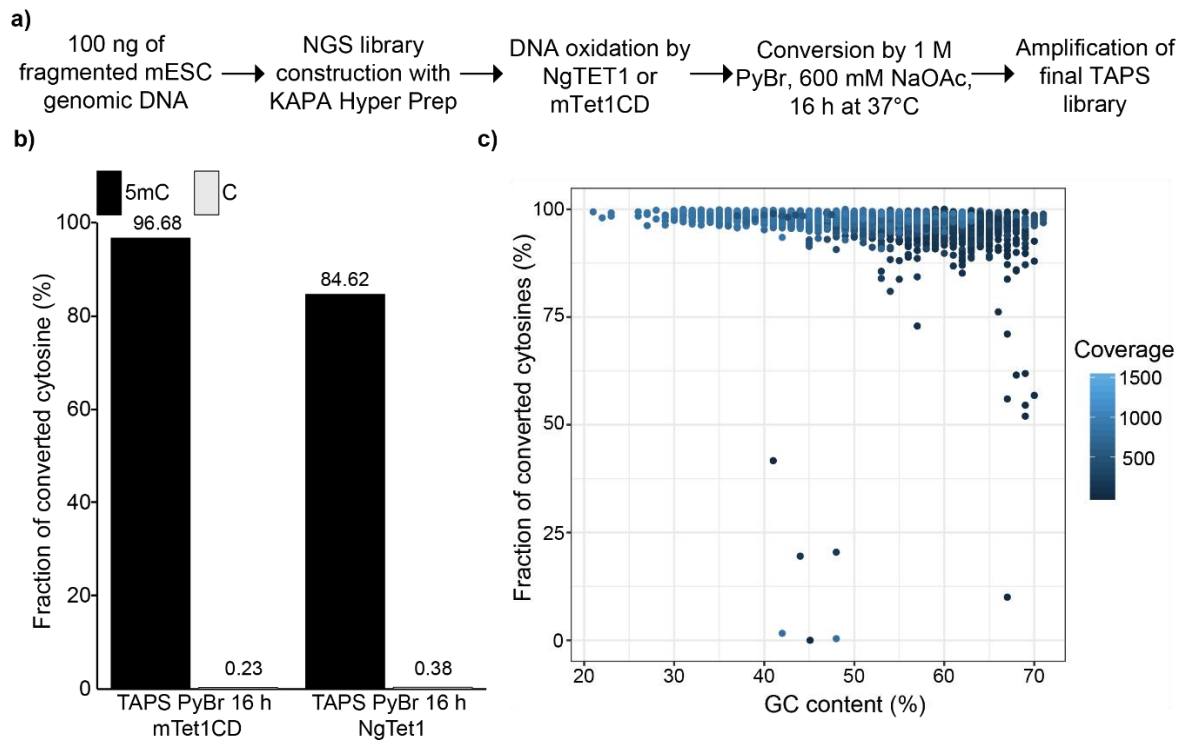


Figure 4-14. Results of test sequencing run to screen different TAPS conditions. (a) General overview of TAPS library preparation according to the optimised protocol. (b) Conversion rate of 5mC in TAPS libraries was estimated based on CpG methylated lambda DNA. False positive rate of TAPS was calculated from unmodified 2kb spike-in. (c) Lack of correlation between 5mC conversion rate of single, methylated CpG in lambda DNA and GC content in 200 bp window around this site (Pearson correlation: -0.16). Sites covered by less than 3 sequencing reads were excluded from analysis.

4.9 TAPS exhibits improved sequencing quality compared to BS-seq and provides reliable whole-genome, base-resolution methylation landscape in mESC

After we established the final conditions for TAPS method, we applied it to the genomic DNA from E14 mESC alongside with the gold standard whole-genome bisulfite sequencing (WGBS) for comprehensive comparison of both methods. TAPS library achieved the highest conversion level for 5mC at CpG context (96.15% in CpG-methylated lambda and 96.9% in synthetic 160mer spike-in) and conversion was less efficient for 5mC in non-CpG context as well as for 5hmC and ranged from 81.2-85.2% (Figure 4-15, a). False positive rate for TAPS conversion was 0.23% with slightly higher

value for unmodified cytosines in CpG compared to those in other sequence contexts (0.35% and 0.18%, respectively, Figure 4-15, b).

As TAPS converts only modified cytosines to thymines, TAPS data have balanced nucleotide composition similar to whole-genome sequencing. This is in contrast to the bisulfite sequencing data where nucleotide composition is seriously affected. Specifically, WGBS data had high proportion (over 40%) of thymine and adenine and extremely low level (below 10%) cytosine and guanine (Figure 4-16).

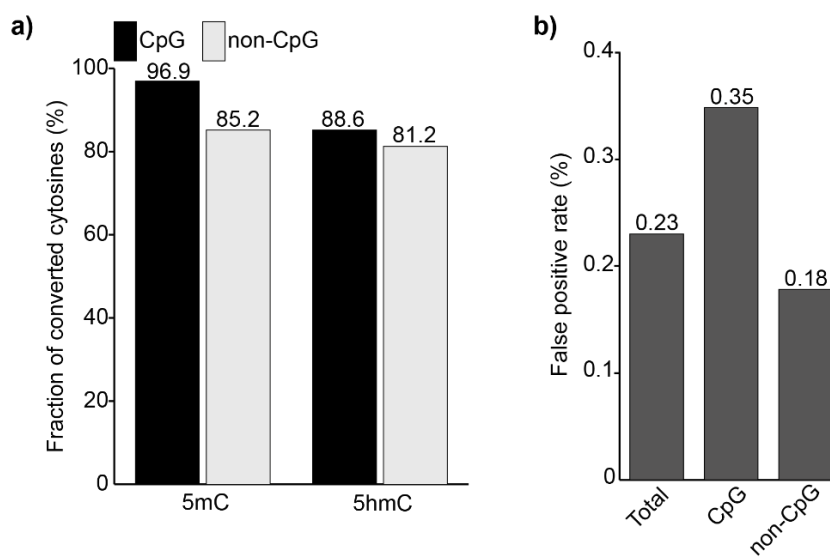


Figure 4-15. Summary of TAPS conversion in whole genome methylome sequencing of E14 mESC. (a) TAPS conversion rate based on synthetic CpN spike-in methylated or hydroxymethylated at known positions. (b) TAPS false-positive rate based on unmodified 2 kb spike-in.

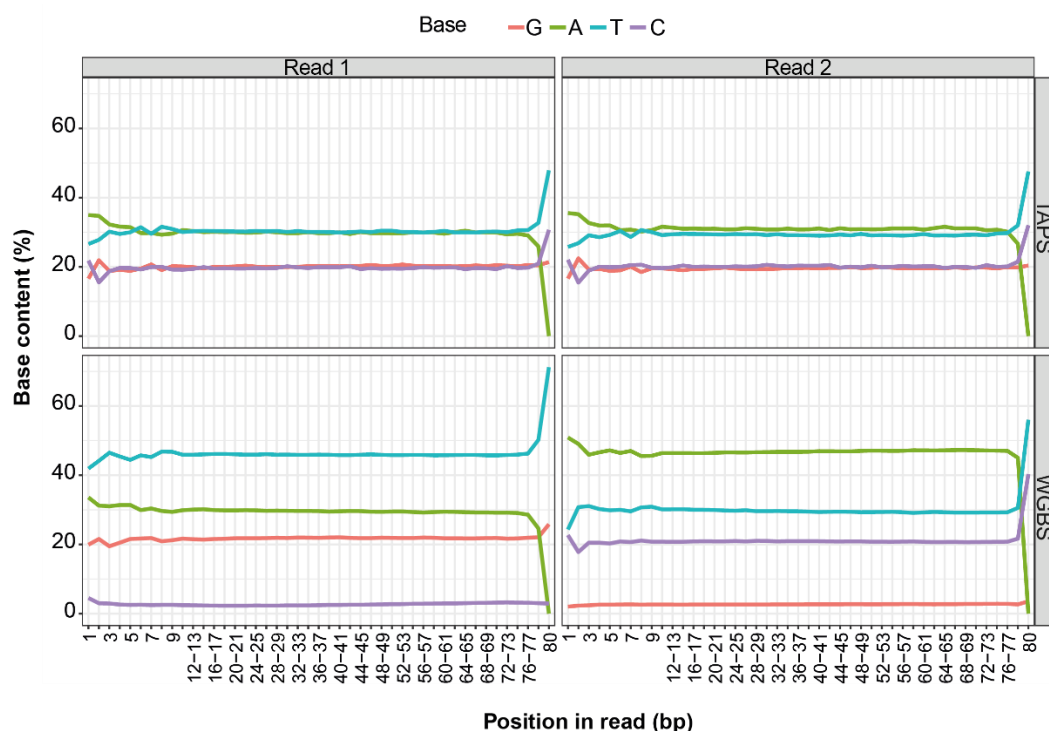


Figure 4-16. Base content in trimmed sequencing reads in TAPS and WGBS. TAPS produces data with complex nucleotide composition similar of whole genome sequencing. In turn, in WGBS most of cytosines in the genome are converted to thymines therefore sequencing data produced by the method have extremely high content of adenines and thymines and very low content of cytosines and guanines.

Improved sequence complexity in TAPS translated into significantly improved Phred sequencing quality scores in cytosine/guanine pairs (Figure 4-17) and improved mapping rate (Figure 4-18). This together with the fact, that WGBS library required supplementation with 15% PhiX to compensate for poor nucleotide complexity, led to higher per-base coverage of TAPS library compared to WGBS at the same sequencing cost (21.8X vs. 13.6X) making TAPS method more cost-effective approach.

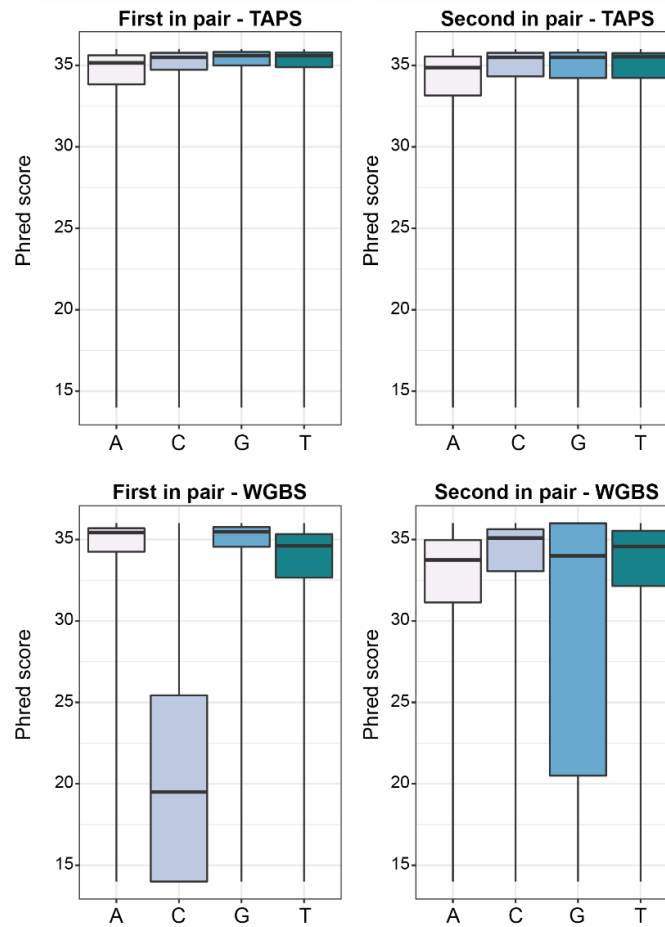


Figure 4-17. Per base sequencing quality scores in first and second reads in trimmed read pairs.

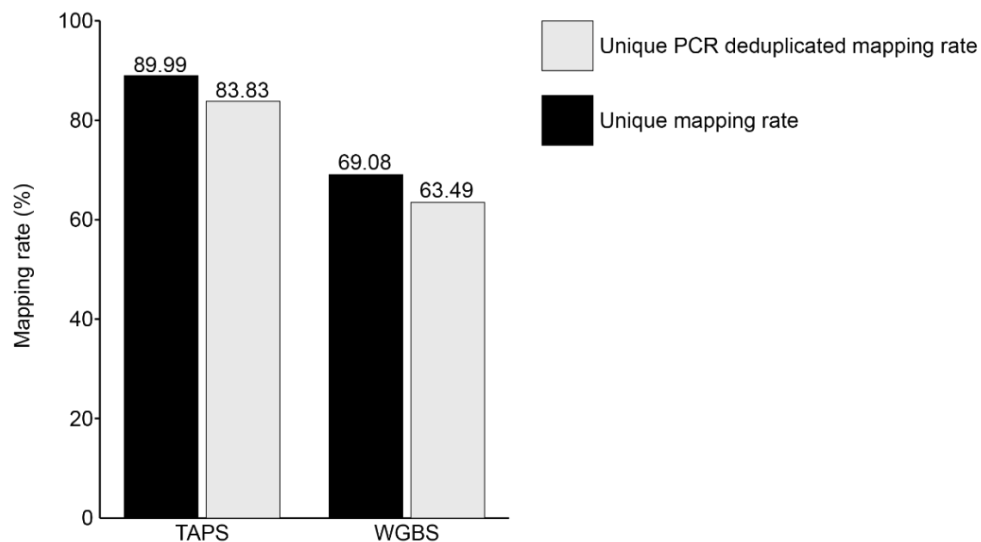


Figure 4-18. Unique mapping rate and unique deduplicated mapping rate in TAPS and WGBS.

After initial comparison, we compared methylation data produced by TAPS and WGBS. In general, TAPS shown higher number of CpG sites covered at least by 3 reads. Out of 42741339 CpG sites in the E14 genome, TAPS detected 88.72% sites compared to

77.80% covered by WGBS. 76.2% (32572816) CpG sites were detected by both methods. Similarly, TAPS had higher number of modified CpG sites (defined by modification value ≥ 0.1) covered by at least 3 reads while both methods consistently detected 29221329 modified CpG (Figure 4-19).

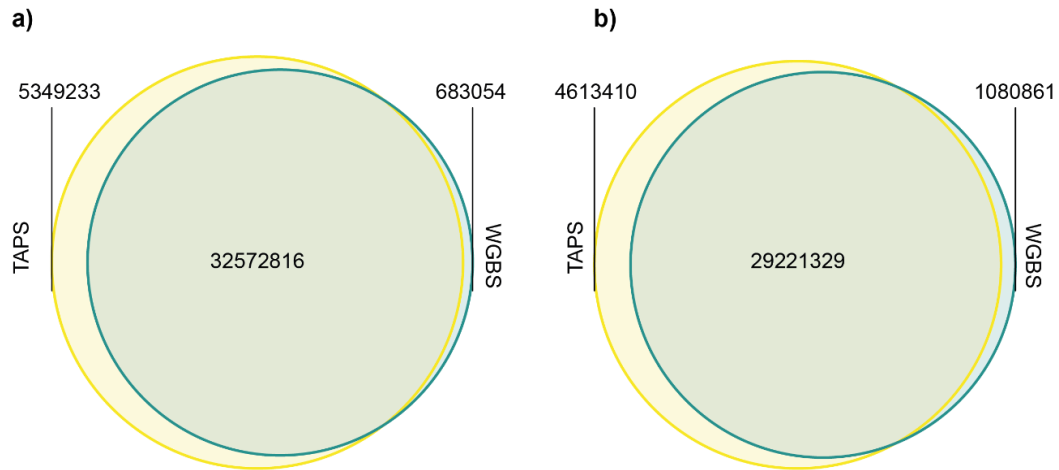


Figure 4-19. (a) CpG sites covered by at least three reads in TAPS, WGBS or both methods. (b) Modified CpG sites (defined by modification level ≥ 0.1) covered by at least three reads in TAPS, WGBS or both methods.

Overall correlation between modification levels of each CpG site covered at least by 3 sequencing reads was high between TAPS and WGBS (Pearson correlation $r = 0.68$) (Figure 4-20). This confirms good agreement between methylation information produced by both methods.

Overall methylation distribution across the chromosomes and genomic features such as CpG islands was very consistent between TAPS and WGBS (Figure 4-21, a and b). However, TAPS exhibited higher and more even coverage in CpG sites even after down-sampling to the coverage similar to WGBS (Figure 4-21, c).

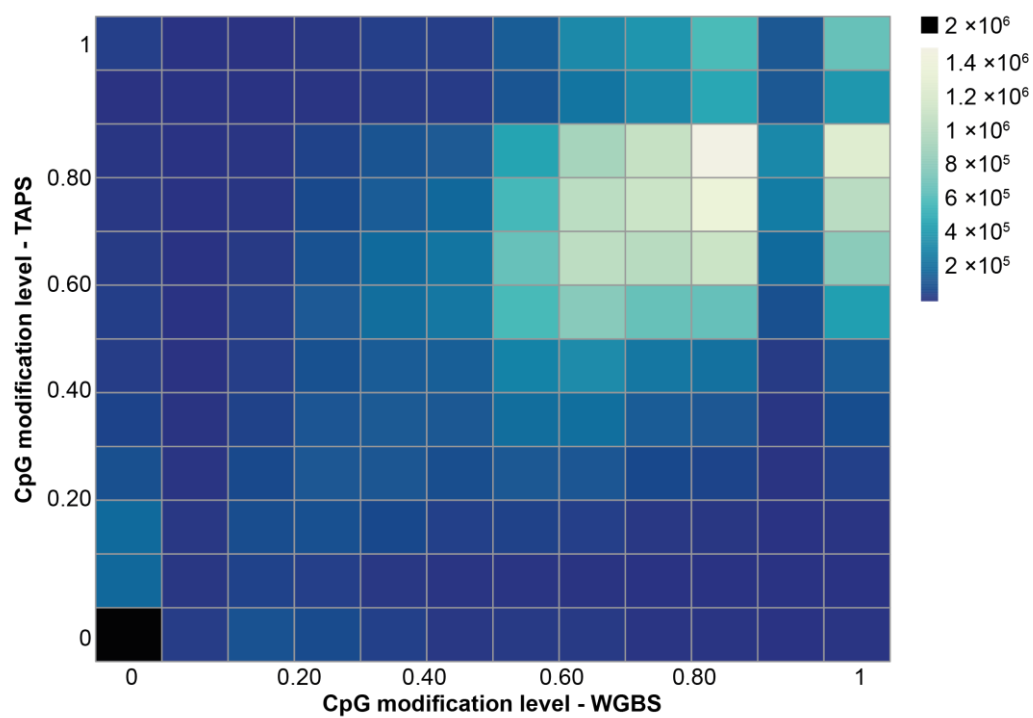


Figure 4-20. Heatmap of the number of CpG sites covered by at least three reads in TAPS and WGBS, grouped by modification level. Bin containing CpG sites detected as unmodified by both methods was excluded from the colour scale to improve contrast and it is coloured in black.

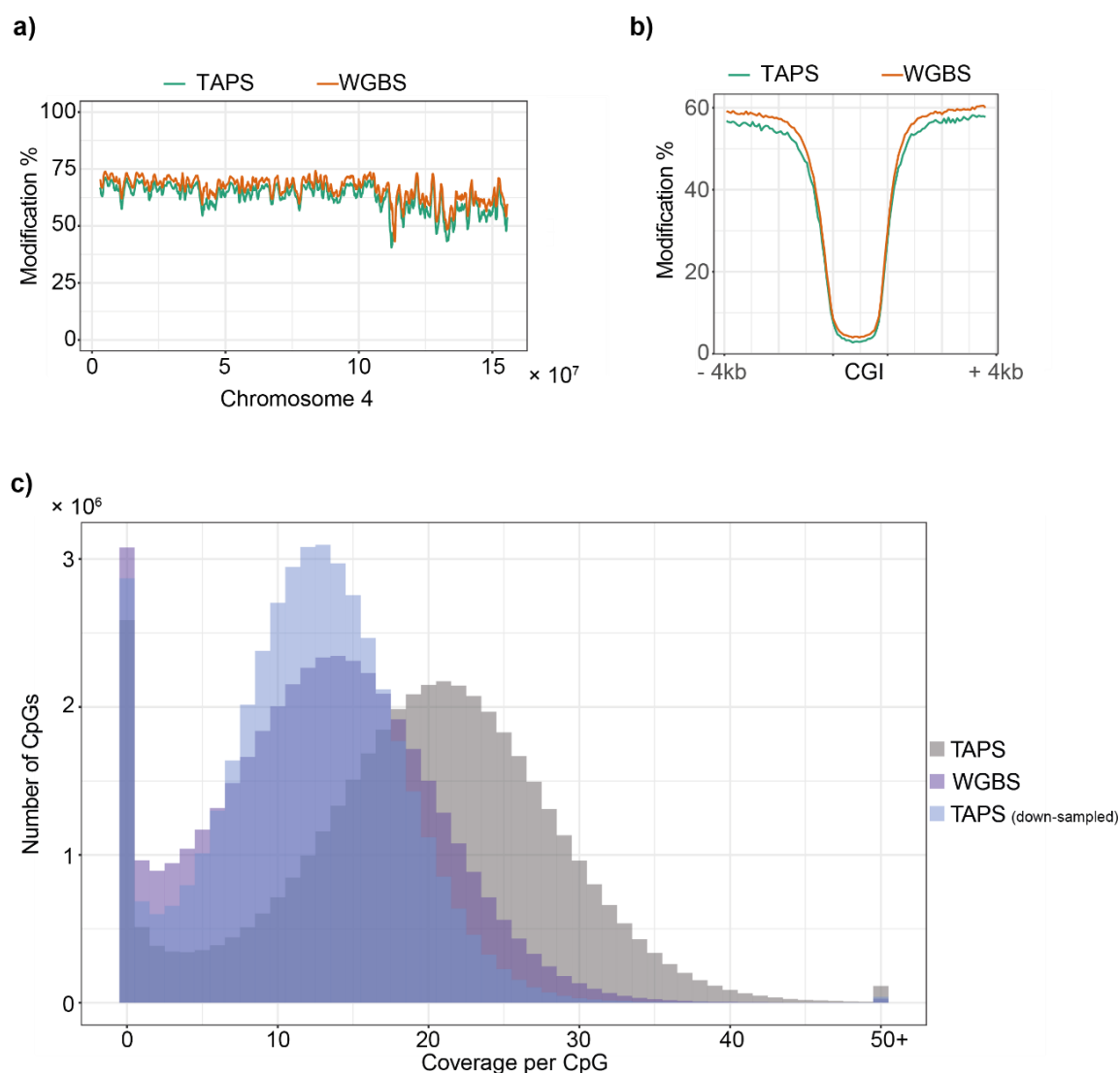


Figure 4-21. (a) Distribution of modification levels (%) in chromosome 4 for TAPS and WGBS. Average CpG modification value was calculated per 100 kb windows along chromosome 4, Gaussian-smoothed (smoothing window size 10). (b) Average modification levels in CGI and surrounding regions. CGI were binned into 20 windows and 4 kb flanking regions were binned into 50 windows. Average modification level was calculated per each bin. (c) Coverage per CpG ($N = 43\,445\,914$ based on mm9 genome) distribution for TAPS and WGBS. For TAPS down-sampled, random reads were selected to matched number of reads for WGBS. Last bin contains positions with coverage equal or higher than 50 $\times$.

We also compared level of DNA damage introduced by the two methods. Extensive fragmentation and denaturation of bisulfite-treated DNA leads to significant sample loss and therefore this method is not well-designed for low input application. Bisulfite-treated DNA becomes single-stranded and can be visualised on the gel only after cooling down. In turn, enzymatic and chemical treatment involved in TAPS protocol retains DNA

double-stranded and DNA can be visualised on the gel immediately after the treatment confirming that our protocol does not denature input DNA. TAPS protocol introduces only minimal double strand breaks and retains DNA fragment over 10 kb long (Figure 4-22, a). Moreover, 5mC conversion was complete in all the samples independently on DNA length (Figure 4-22, b-c). This finding will facilitate application of TAPS for more challenging applications such as low-input samples or long-read sequencing.

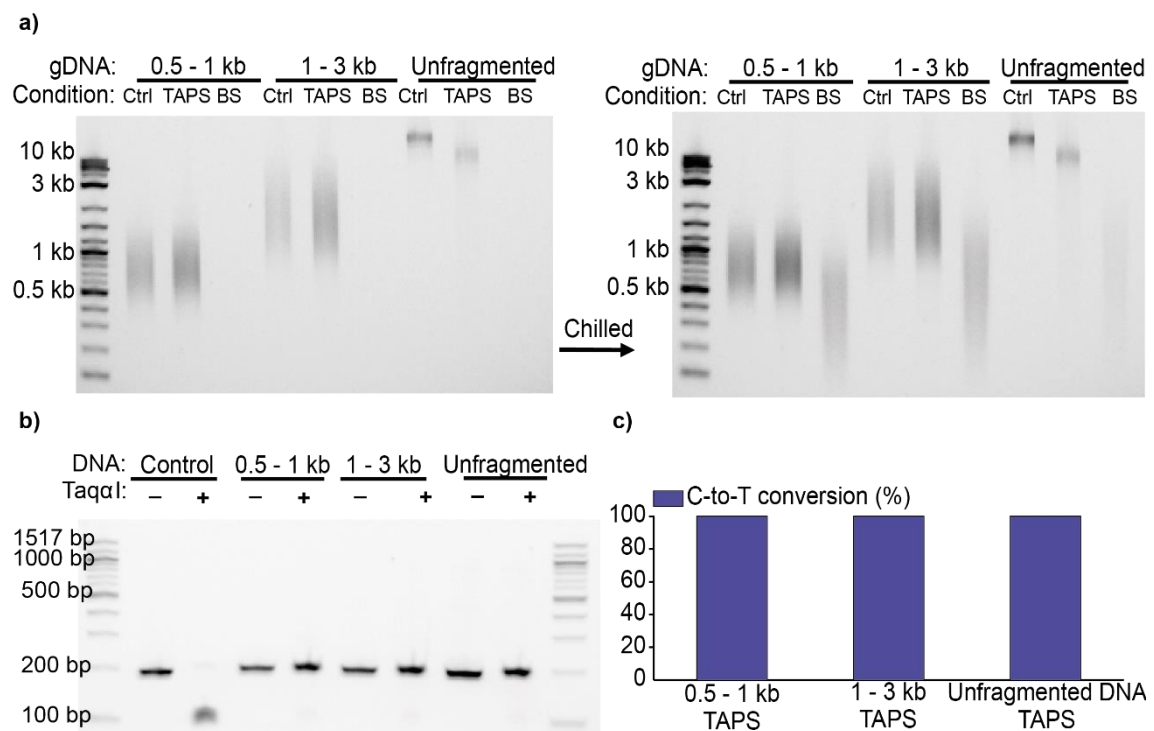


Figure 4-22. TAPS on mESC genomic DNA of various fragment lengths. (a) Comparison of genomic DNA samples oxidised with mTet1CD and reduced with pyridine borane or treated with bisulfite before (left panel) and after (right panel) cooling down on ice. DNA after TAPS remained double stranded and could be directly visualized on the gel. Bisulfite treatment caused more damage and fragmentation to the samples and DNA became single stranded and could be visualized only after being chilled on ice. Experiment was performed once. (b) Agarose gel images of the TaqI digestion assay confirming complete 5mC-to-T conversion in all samples regardless of DNA fragment length. 194 bp model sequence from the lambda genome was PCR amplified after TAPS and digested with TaqI enzyme. The PCR product amplified from unconverted sample could be cleaved, whereas products amplified on TAPS treated samples stayed intact, suggesting loss of restriction site and hence complete C-to-T transition. Experiment was performed once. (c) The C-to-T conversion percentage was estimated by gel band quantification as 100% for all DNA fragment lengths tested.

4.10 Discussion

In this chapter we described the development of a new method for base-resolution mapping of 5mC and 5hmC, TAPS, and its comparison with the gold standard – bisulfite sequencing. In this section I would like to discuss some technical aspects of TAPS method and its development.

4.10.1 Production of recombinant TET enzymes

Complete oxidation of 5mC and 5hmC to 5caC by TET is crucial for successful TAPS conversion. Although both pyridine borane and pic-borane were shown to work both on 5fC and 5caC, 5fC may be less stable in chemical reactions. Booth *et al.* already demonstrated chemical conversion of 5fC to 5hmC (Booth *et al.*, 2014). Even low level of an uncontrolled conversion of 5fC to 5hmC could lead to increased false negative rate in TAPS and therefore 5caC remains to be the most optimal substrate for borane reduction. Moreover, an ideal enzyme for TAPS should be free from nuclease contamination and stable during storage to facilitate its application to the method. Recombinant TET enzymes from various organisms including mouse (Yu *et al.*, 2012), human (L. Hu *et al.*, 2013; Hu *et al.*, 2015; Liu *et al.*, 2017), *Naegleria guberi* (Hashimoto *et al.*, 2014, 2015; Pais *et al.*, 2015; Tamanaha *et al.*, 2016) and *Coprinopsis cinerea* (Zhang *et al.*, 2014) have been already described in the literature using various recombinant expression systems including protein expression in bacteria, insect or mammalian cells.

We initially selected NgTET1 as a protein easy to produce at high quantity and good activity (Hashimoto *et al.*, 2014, 2015; Pais *et al.*, 2015; Tamanaha *et al.*, 2016). The major drawback of NgTET1 protein is its oxidative activity on thymine oxidation but this off-target effect was estimated as low by other research (Pais *et al.*, 2015). Contrary,

thymine oxidation products generated particularly high signal in MALDI assays in our TET activity tests. However, MALDI has limited quantitative potential (Bucknall, Fung and Duncan, 2002; Szájli, Fehér and Medzihradszky, 2008) and thus cannot be used as a proof of high level of T oxidation caused by NgTET1 enzyme produced in our lab. Relatively high signal for thymine oxidation products could be both due to short length of tested oligo and the presence of multiple thymines in the sequence. On the other hand, we cannot exclude possibility that low level of off-target activity of NgTET1 could contribute to the problems with low conversion rate and high false positive rate observed at the initial stages of method development. In fact, TAPS performance improved significantly after implementing mTet1CD to the method. Although mTet1CD can be produced at good quality and activity, its production is relatively expensive and difficult to scale up, which significantly reduces TAPS method's potential to be applied to the large number of samples.

Systematic screening of recombinant TET enzymes and optimisation of the most promising enzymes could be good and efficient strategy for finding highly active and easier to express and purify TET enzyme for TAPS in the future. Some approaches could be applied to make TET production more robust. They include:

- Optimisation of mTET1CD expression in mammalian or insect cells by viral infections
- Optimisation of mTET1CD or other TET proteins for expression in bacteria
- Expressing of non-mammalian TET enzymes and its optimisation for TAPS

Briefly, expressing TET enzymes in mammalian or insect cells by viral infection instead of transient transfection could potentially improve reproducibility of protein yield between expression batches and simplify expression protocol. On the other hand, based on small scale experiments we found that optimisation of viral titer to obtain reasonable

protein yield without affecting overall protein stability and inducing degradation was very difficult. Therefore, we did not consider this approach as more advantageous than currently used transient expression in mammalian cells.

More beneficial approach would be expression of TET enzymes in bacteria. Some non-mammalian TET enzymes (e.g. NgTET1 and ccTET1) can be easily expressed in *E. coli* cells, however, they may be less active than their mammalian counterparts and they may pose additional, undesired activities (Zhang *et al.*, 2014; Pais *et al.*, 2015). In turn, mammalian TET enzymes are difficult to produce in bacteria due to the presence of an unstructured region within a catalytic domain that affects their stability during expression and purification. This can be overcome by mutating catalytic domain to remove problematic regions as demonstrated previously (L. Hu *et al.*, 2013; Hore *et al.*, 2016). In our lab, we managed to produce modified hTET2 in *E. coli*, however we found that mTet1CD is still better and more robust oxidation enzyme.

To summarise, although produced mTet1CD is good and active enzyme for TAPS applications, further improvements may be required to optimise its production to the large-scale level. In this short discussion I proposed some future directions for improvement of recombinant TET production for TAPS.

4.10.2 Methylation analysis by TAPS

We demonstrated that TAPS can specifically and efficiently detect methylated cytosines and has unique characteristics compared to WGBS. Since TAPS protocol does not involve denaturing, DNA damaging conditions and retains double-stranded, non-fragmented DNA it offers new opportunities to develop derivative methods or to apply TAPS to challenging samples.

The clear advantage of TAPS over BS-seq is its ability to introduce base change at modified sites. Therefore, TAPS protocol converts only a small fraction of cytosines, and sequencing data can be still used for additional purposes such as mutation analysis. However, as C-to-T mutations are the most common in the genome (Alexandrov *et al.*, 2013), additional computational approaches need to be developed to ensure high-confidence detection of this kind of mutations and distinguishing them from methylation. Alternatively, removing the common SNPs from a dataset prior to downstream analysis is highly recommended for more specific methylation calling.

Although TAPS method is a great technological improvement, one limitation of the current method is lower conversion on 5hmC and at non-CpG context compared to 5mC. This can be explained by lower activity of TET enzyme in non-CpG context (L. Hu *et al.*, 2013) or when oxidation substrate is 5hmC (Hu *et al.*, 2015). Lower conversion underestimates modification values in 5hmC and outside CpG sites and can explain slightly lower modification levels measured by TAPS compared to BS-seq. Low non-CpG conversion can be neglected in most of cases as the majority of methylated cytosines in mammalian cells occur in CpG context. However, this can be a problem in species or cell types with high level of methylation outside the CpG context e.g. plants (Law and Jacobsen, 2010; Zhang and Pradhan, 2014). Lower conversion in 5hmC will only weakly affect TAPS signal which is a sum of 5mC and 5hmC as 5hmC abundance is much lower than 5mC in somatic tissues and will comprise only a small proportion of sites (Yu *et al.*, 2012). It is also important to highlight that although TAPS performance on 5hmC conversion was not optimal, most of currently available base-resolution methods cannot detect 5hmC at higher sensitivity (Yu *et al.*, 2012). However, 5hmC TAPS conversion requires further improvements to achieve highly sensitive and accurate detection of hydroxymethylated sites.

4.10.3 Concluding remarks

We demonstrated that TAPS can specifically and accurately detect 5mC and 5hmC in the genome and in many aspects outperforms bisulfite sequencing. In particular, TAPS directly introduces a base change of 5mC and 5hmC and therefore solves bisulfite sequencing technical issues associated with low sequence complexity after conversion. This is further translated into improvements in mapping rate and better sequencing quality. Non-damaging character of TAPS reaction opens new possibilities to its application to cell-free DNA analysis, single-cell sequencing and long read sequencing.

4.11 Materials and Methods

4.11.1 Preparation of DNA oligonucleotides for MALDI

Unmodified DNA oligonucleotides and oligonucleotides with 5mC modification were purchased from IDT. Sequences of DNA oligonucleotides are listed in Table 4-2.

5mC double-stranded DNA oligonucleotide was generated by annealing 5mC_11mer and 17mer oligos.

The DNA oligo with 5caC modification was synthesized using Expedite 8900 Nucleic Acid Synthesis System (Thermo Fisher) with standard phosphoramidites (Sigma) and 5-Carboxy-dC-CE Phosphoramidite (Glen Research). Subsequent deprotection and purification were carried out with Glen-Pak Cartridges (Glen Research) according to the manufacturer's instructions. Purified oligonucleotides were characterized by Voyager-DE MALDI-TOF Biospectrometry Workstation.

Table 4-2. DNA oligonucleotides used for preparation of model DNA oligos for MALDI.

DNA oligonucleotide	Sequence (5' → 3')
5mC_11mer	GTCGAC/5mC/GGATC
17mer	TTGGATCCGGTCGACTT
C_11mer	GTCGACCGGATC

4.11.2 Preparation of spike-in DNA controls for TAPS

CpG-methylated lambda DNA

1 µg of unmethylated lambda DNA (Promega) was methylated in the presence of 0.64 mM SAM and 0.8 U/µL M. SssI enzyme (NEB) in Mg²⁺-free buffer (10 mM Tris-Cl pH 8.0, 50 mM NaCl, and 10 mM EDTA) for 2 h at 37°C in a 50 µL reaction. Then, 0.5 µL of M. SssI enzyme and 1 µL of SAM were added and the reaction was incubated for an additional 2 h at 37°C. Methylated DNA was subsequently purified on 1× AMPure XP beads. To assure complete methylation, whole procedure was repeated a second time in NEBuffer 2. DNA methylation was then validated with HpaII digestion assay. 50 ng of methylated and unmethylated DNA were digested in a 10 µL reaction with 2 U of HpaII enzyme (NEB) in CutSmart buffer (NEB) for 1 h at 37°C. Digestion products were run on a 1% agarose gel together with undigested lambda DNA control. Unmethylated lambda DNA was digested after the assay whereas methylated lambda DNA remained intact confirming complete and successful CpG methylation. The sequence of lambda DNA can be found in GenBank - EMBL Accession Number: J02459.

2kb spike-in control

2kb spike-in control was PCR amplified from pNIC28-Bsa4 plasmid (Addgene, cat. no. 26103) in the reaction containing 1 ng DNA template, 0.5 µM primers, 1 U Phusion High-Fidelity DNA Polymerase (Thermo Fisher). The amplicon was purified on Zymo-Spin columns (Zymo Research) after PCR amplification. PCR primers are listed in Table 4-3. Full sequences of 2kb spike-in are listed in Appendix 4 Supplementary Table 4-1.

Table 4-3. PCR primers for 2kb spike-ins amplification.

Primer	Sequence (5'→3')
2kb_Foward	CACAGATGTCTGCCTGTTCA
2kb_Reverse	AGGGTGGTGAATGTGAAACC

Short synthetic spike-in control

A synthetic oligo with N5mCNN and N5hmCNN sequences was produced by the annealing and extension method. Briefly, 10 μ M N5mCNN and N5hmCNN oligos (IDT) were annealed in the annealing buffer containing 5 mM Tris-Cl (pH 7.5), 5 mM MgCl₂, and 50 mM NaCl. Extension was performed in the NEB buffer 2 with 0.4 mM dNTPs (dATP/dGTP/dTTP/dCTP) and 5 U of Klenov Polymerase (NEB) for 1 hour at 37°C. After the reaction, the spike-in control was purified on Zymo-Spin IC column (Zymo Research). Oligo sequences are listed in Table 4-4.

Table 4-4. Modified DNA oligonucleotides for preparation synthetic oligo control with N5mCNN and N5hmCNN.

Oligo	Template sequence (5' to 3')
N5mCNN	GAAGATGCAGAAGACAGGAAGGATGAAACACTCAGGCGCA CGCTGGCATNmCNNGACAAACCACAAGAACAGGCTAGTGA GAATGAAGGGA
N5hmCNN	CCAACTCTGAAACCCACCAACGCCAACATCCACCACACAAC CCAAGATNhmCNNGACCATCTTACAAACATATCCCTTCATTC TCACTAGCC

4.11.3 Preparation of 222 bp model DNA

The 222 bp model DNA containing five CpG sites was generated from lambda DNA (Thermo Fisher) by PCR amplification using Taq DNA Polymerase (NEB) and purified by AMPure XP beads (Beckman Coulter). Primers sequences are as follows: FW-5'-CCTGATGAAACAAGCATGTC-3', RV-5'-CAUACTCACUTCCCCACUT-3'. The uracil base in the reverse strand of PCR product was removed by USER enzyme (NEB). 100 ng of purified PCR product was then methylated in 20 μ l solution containing 1X NEBuffer 2, 0.64 mM S-adenosylmethionine and 20 U M. SssI CpG Methyltransferase

(NEB) for 2 h at 37°C, followed by 20 min heat inactivation at 65°C. The methylated 222 bp model DNA was purified by AMPure XP beads.

4.11.4 DNA digestion and HPLC–MS/MS analysis

DNA samples were digested with 2 U of Nuclease P1 (Sigma-Aldrich) and 10 nM deaminase inhibitor erythro-9-amino- β -hexyl- α -methyl-9H-purine-9-ethanol hydrochloride (Sigma-Aldrich). After overnight incubation at 37 °C, the samples were further treated with 6 U of alkaline phosphatase (Sigma-Aldrich) and 0.5 U of phosphodiesterase I (Sigma-Aldrich) for 3 h at 37 °C. The digested DNA solution was filtered with Amicon Ultra-0.5 ml 10K centrifugal filters (Merck Millipore) to remove the proteins and subjected to HPLC–MS/MS analysis. The HPLC–MS/MS analysis was carried out with 1290 Infinity LC Systems (Agilent) coupled with a 6495B Triple Quadrupole Mass Spectrometer (Agilent). A ZORBAX Eclipse Plus C18 column (2.1 $\times$ 150 mm², 1.8- μ m, Agilent) was used. The column temperature was maintained at 40 °C, and the solvent system was water containing 10 mM ammonium acetate (pH 6.0, solvent A) and water–acetonitrile (60/40, v/v, solvent B) with 0.4 ml min^{–1} flow rate. The gradient was: 0–5 min, 0% solvent B; 5–8 min, 0–5.63% solvent B; 8–9 min, 5.63% solvent B; 9–16 min, 5.63–13.66% solvent B; 16–17 min, 13.66–100% solvent B; 17–21 min, 100% solvent B; 21–24.3 min, 100–0% solvent B; 24.3–25 min, 0% solvent B. The dynamic multiple reaction monitoring mode of the mass spectrometer was used for quantification. The source-dependent parameters were as follows: gas temperature 230 °C, gas flow 14 l min^{–1}, nebulizer 40 p.s.i., sheath gas temperature 400 °C, sheath gas flow 11 l min^{–1}, capillary voltage 1,500 V in the positive ion mode, nozzle voltage 0 V, high-pressure RF 110 V and low pressure RF 80 V, both in the positive ion mode.

The fragmentor voltage was 380 V for all compounds, while other compound-dependent parameters are summarized in Appendix 4 Supplementary Table 4-2.

4.11.5 Expression and purification of mTet1CD

The mTet1CD catalytic domain (NM_001253857.2, 4371-6392) with N-terminal Flag-tag was cloned into pcDNA3-Flag between the *KpnI* and *BamHI* restriction sites (Yu *et al.*, 2012). mTET1CD protein details are summarised in Appendix 4 Supplementary Table 4-3. For protein expression, 1 mg plasmid was transfected into 1 l of Expi293F cell culture at density 1×10^6 cells ml⁻¹ and cells were grown for 48 h at 37 °C, 170 r.p.m. and 5% CO₂. Subsequently, cells were harvested by centrifugation, re-suspended in the lysis buffer containing 50 mM Tris–Cl pH 7.5, 500 mM NaCl, 1× cOmplete Protease Inhibitor Cocktail, 1 mM PMSF, 1% Triton X-100 and incubated on ice for 20 min. Cell lysate was then clarified by centrifugation for 30 min at 30,000*g* and 4°C. Collected supernatant was purified on ANTI-FLAG M2 Affinity Gel and pure protein was eluted with buffer containing 20 mM HEPES pH 8.0, 150 mM NaCl, 0.1 mg ml⁻¹ 3× Flag peptide, 1× cOmplete Protease Inhibitor Cocktail, 1 mM PMSF. Collected fractions were concentrated and buffer exchanged to the final buffer containing 20 mM HEPES pH 8.0, 150 mM NaCl and 1 mM dithiothreitol. Concentrated protein was mixed with glycerol (30% v/v), frozen in liquid nitrogen and aliquots were stored at –80°C.

4.11.6 TET activity assay

For NgTET1 activity assay 5 µM 5mC 11mer oligonucleotide was oxidised in 20 µL reaction containing 50 mM MOPS buffer (pH 6.9), 100 µM ammonium iron(II) sulfate, 1 mM α-ketoglutarate, 2 mM ascorbic acid, 1 mM dithiothreitol, 50 mM NaCl and 5 µM NgTET1 for 1 h at 37°C. After that, 4 U of Proteinase K (New England Biolabs) were added to the reaction mixture and incubated for 30 min at 37°C.

For mTet1CD activity assay 5 μ M 5mC 11mer oligonucleotide were incubated in a 20 μ l reaction containing 50 mM HEPES buffer (pH 8.0), 100 μ M ammonium iron(II) sulfate, 1 mM α -ketoglutarate, 2 mM ascorbic acid, 1 mM dithiothreitol, 100 mM NaCl, 1.2 mM ATP and 4 μ M mTet1CD for 80 min at 37°C. After that, 0.8 U of Proteinase K (New England Biolabs) were added to the reaction mixture and incubated for 1 h at 50°C. Oxidation products were purified by Bio-Spin P-30 Gel Column (Bio-Rad) and analysed by Voyager-DE MALDI-TOF Biospectrometry Workstation.

4.11.7 NgTET1 oxidation

Genomic DNA (up to 500 ng) was incubated in 50 μ l solution containing 50 mM MOPS buffer (pH 6.9), 100 μ M ammonium iron(II) sulfate, 1 mM α -ketoglutarate, 2 mM ascorbic acid, 1 mM dithiothreitol, 50 mM NaCl and 5 μ M NgTET1 for 1 h at 37°C. After that, 4 U of Proteinase K (New England Biolabs) were added to the reaction mixture and incubated for 30 min at 37°C. The product was cleaned up on 1.8 $\times$ AMPure XP beads following the manufacturer's instruction.

4.11.8 mTet1CD oxidation

Genomic DNA (up to 200 ng) was incubated in a 50 μ l reaction containing 50 mM HEPES buffer (pH 8.0), 100 μ M ammonium iron(II) sulfate, 1 mM α -ketoglutarate, 2 mM ascorbic acid, 1 mM dithiothreitol, 100 mM NaCl, 1.2 mM ATP and 4 μ M mTet1CD for 80 min at 37 °C. After that, 0.8 U of Proteinase K (New England Biolabs) were added to the reaction mixture and incubated for 1 h at 50°C. The product was cleaned up on Bio-Spin P-30 Gel Column (Bio-Rad) and 1.8 $\times$ AMPure XP beads following the manufacturer's instruction.

4.11.9 Validation of TAPS conversion with Taq α I assay

The converted DNA sample was PCR amplified by Taq DNA Polymerase (New England Biolabs) with the corresponding primers according to manufacturer's protocol. The PCR product was incubated with 4 units of Taq α I restriction enzyme (New England Biolabs) in 1 $\times$ CutSmart buffer (New England Biolabs) for 30 min at 65 °C and visualized by 2% agarose gel electrophoresis. Primer sequences are listed in Table 4-5. Detailed sequences of control sequences are listed in Appendix 4 Supplementary Table 4-4.

Table 4-5. Sequences of PCR primers for amplification of control sequences for Taq α I assay.

ID	Forward primer (5' $\rightarrow$ 3')	Reverse primer (5' $\rightarrow$ 3')
222mer	CCTGATGAAACAAGCATGTC	CATTACTCACTTCCCCACTT
194mer	GCTGGGGAAGTACAGGCT	AGAACCAGAACTCAAAGTGTAC
216mer	GTCAGGTTACCAACTACT	AAAAGTACCTCTTCCAG

4.11.10 TAPS on model DNA

500 ng of model 222bp CpG-methylated DNA were oxidised by NgTET1 as described above. Oxidised DNA was further reduced by pic-borane and amplified by Taq DNA Polymerase (New England Biolabs) or KAPA Hifi Uracil Plus Polymerase. Conversion was tested by Taq α I digestion assay as described above. The PCR product was purified by either Exonuclease I (New England Biolabs) and Shrimp Alkaline Phosphatase (New England Biolabs) or Zymo-Spin columns and then processed for Sanger sequencing.

4.11.11 Preparation of Illumina multiplexing adapters

Adapter oligos (5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCT-3', 5'-/5Phos/GATCGGAAGAGCACACGTCT-3') were obtained from IDT with HPLC purification. Adapter oligos were annealed together in a 50 μ L reaction containing 15 μ M of each oligo, 10 mM Tris-Cl (pH = 8.0), 0.1 mM EDTA (pH = 8.0) and 50 mM NaCl with the following program: 2 min at 95°C, 140 cycles of 20 sec at 95°C (decrease

temperature 0.5°C every cycle) and hold at 4°C. Annealed 15 µM Illumina multiplexing adapters were then aliquoted into small single-use vials and stored at -80°C.

4.11.12 mESC culture and genomic DNA preparation

E14 mESC (gift from S. Kriaucionis) were cultured on gelatin-coated plates in DMEM (Invitrogen) supplemented with 15% FBS (Gibco), 2 mM L-glutamine (Gibco), 1% non-essential amino acids (Gibco), 1% penicillin/streptavidin (Gibco), 0.1 mM β-mercaptoethanol (Sigma), 1,000 units ml⁻¹ leukaemia inhibitory factor (Millipore), 1 µM PD0325901 (Stemgent) and 3 µM CHIR99021 (Stemgent). Cultures were maintained at 37°C and 5% CO₂ and passaged every 2 days.

For isolation of genomic DNA, cells were harvested by centrifugation for 5 min at 1,000g and room temperature. DNA was extracted with Quick-DNA Plus kit (Zymo Research) according to the manufacturer's protocol.

4.11.13 Pic-borane TAPS (in MeOH) sequencing of mESC genomic DNA

mESC gDNA was spiked-in with 0.1% CpG-methylated lambda DNA and 0.01% unmodified 2kb-3. DNA was then sonicated with a Covaris M220 instrument and size-selected to 200-400 bp using AMPure XP beads. 1 µg of size-selected DNA was then double-oxidised with NgTET1 as described above. Oxidised DNA was then used to End-repair and A-tailing reaction and ligation of Illumina adapter (NextFlex) with prepared with KAPA HyperPrep kit (Kapa Biosystems) according to manufacturer's protocol. Oxidised DNA was then reduced in a 50 µL reaction containing 50% MeOH (v/v), 2.5 M pic-BH₃ and 300 mM NaOAc pH = 4.3 for 8 h at 60°C at 850 r.p.m. in an Eppendorf ThermoMixer and purified by Zymo-Spin IC columns. Final library was amplified with KAPA Hifi Uracil Plus Polymerase and purified on 1× AMPure XP beads. The whole-genome TAPS library was paired-end 80 bp sequenced on a NextSeq 500 sequencer

(Illumina) using one NextSeq High Output kit with 1% PhiX control library spike-in. 50 ng of gDNA before TAPS, after NgTET1 oxidation and pic-BH₃ reduction were analysed by HPLC-MS/MS as described above.

4.11.14 TAPS optimisation with pic-BH₃

500 ng fragmented and size-selected mESC gDNA (200-400 bp) spiked-in with 0.5% CpG methylated lambda DNA were double-oxidised by NgTET1 as described above. 2-Picoline-borane (100 mg) (pic-borane, Sigma-Aldrich) was dissolved in 187 µl of MeOH or DMSO to give around a 3.26 M solution. In each reaction, 50 ng of oxidised DNA was reduced in a 50 µL reaction containing 25 µl of pic-borane solution with or without buffer solution. Reactions were incubated at different temperatures and for different time periods at 850 r.p.m. in an Eppendorf ThermoMixer and purified by Zymo-Spin IC column. TAPS conversion was tested by TaqαI assay on 194mer and/or 216mer control sequences as described above. 50 ng of DNA oxidised by NgTET1 and after reduction under certain conditions were analysed by HPLC-MS/MS as described above. Summary of tested conditions is in Table 4-6.

Table 4-6. Summary of TAPS condition tested during conditions optimisation with pic-BH₃.

Condition	Solvent	Buffer	Incubation	HPLC-MS/MS
Pic-BH₃-1	MeOH	0.3 M NaOAc pH = 5.2	2 h at 60°	No
Pic-BH₃-2	MeOH	0.3 M NaOAc pH = 5.2	5 h at 60°	No
Pic-BH₃-3	DMSO	0.3 M NaOAc pH = 5.2	2 h at 60°	No
Pic-BH₃-4	DMSO	0.3 M NaOAc pH = 5.2	2 h at 70°	No
Pic-BH₃-4	DMSO	0.3 M NaOAc pH = 5.2	5 h at 60°	No
Pic-BH₃-5	DMSO	0.3 M NaOAc pH = 5.2	3 h at 70°	Yes
Pic-BH₃-6	DMSO	0.3 M NaOAc pH = 5.2	4 h at 70°	Yes
Pic-BH₃-7	DMSO	0.3 M NaOAc pH = 5.2	5 h at 70°	Yes
Pic-BH₃-8	DMSO	0.3 M NaOAc pH = 5.2	8 h at 70°	Yes
Pic-BH₃-9	DMSO	0.3 M MES pH = 4.0	5 h at 70°	No
Pic-BH₃-10	DMSO	0.3 M HEPES pH = 7.0	5 h at 70°	No

4.11.15 Pic-borane TAPS sequencing of mESC genomic DNA

mESC gDNA was spiked-in with 0.1% CpG-methylated lambda DNA and 0.01% unmodified 2kb-3. DNA was then sonicated with a Covaris M220 instrument and size-selected to 200-400 bp using AMPure XP beads. 1 µg of size-selected DNA was then double-oxidised with NgTET1 as described above. Oxidised DNA was then used to End-repair and A-tailing reaction and ligation of Illumina adapter (NextFlex) with prepared with KAPA HyperPrep kit (Kapa Biosystems) according to manufacturer's protocol. 100 mg of 2-picoline-borane (pic-borane, Sigma-Aldrich) was dissolved in 187 µL of DMSO to give around 3.26 M solution. For each reaction, 25 µL of pic-borane solution and 5 µL of 3 M sodium acetate solution (pH 5.2, Thermo Fisher) were added into 20 µL of DNA sample and incubated for 3 hours at 70°C. The product was purified by Zymo-Spin IC following the manufacturer's instructions. Final library was amplified with KAPA Hifi Uracil Plus Polymerase and purified on 1× AMPure XP beads. The whole-genome TAPS sequencing library was paired-end 80 bp sequenced on a NextSeq 500 sequencer (Illumina) using one NextSeq High Output kit with 1% PhiX control library spike-in. 50 ng of gDNA before TAPS, after NgTET1 oxidation and pic-BH₃ reduction were analysed by HPLC-MS/MS as described above.

4.11.16 Low-input whole-genome TAPS (pic-BH₃) with ssDNA library kit

mESC gDNA was purified and fragmented as described above. Briefly, samples containing 100 ng, 10 ng, 1 ng, 100 pg, and 10 pg of mESC gDNA were oxidized with NgTET1 once and reduced with pic-borane as described above. Sequencing libraries were prepared with Accel-NGS Methyl-Seq DNA Library Kit (Swift Biosciences) according to manufacturer's protocol. Final libraries were amplified with KAPA Hifi

Uracil Plus Polymerase for 6 cycles (100 ng), 9 cycles (10 ng), 13 cycles (1 ng), 16 cycles (100 pg), and 21 cycles (10 pg) and cleaned-up on 0.85× AMPure XP beads.

4.11.17 Cell-free DNA TAPS (pic-BH₃)

Cell-free TAPS libraries were prepared from 10 ng and 1 ng of cfDNA sample. Briefly, samples were oxidized with NgTET1 once and reduced with pic-borane as described above. Sequencing libraries were prepared with Accel-NGS Methyl-Seq DNA Library Kit (Swift Biosciences) according to manufacturer's protocol. Final libraries were amplified with KAPA Hifi Uracil Plus Polymerase for 9 cycles (10 ng) and 13 cycles (1 ng) and cleaned-up on 0.85× AMPure XP beads.

4.11.18 TAPS with pyridine borane – initial conditions

For whole-genome TAPS, mESC genomic DNA was spiked with 0.5% of methylated lambda DNA and 0.025% of unmodified 2 kb spike-in control. DNA samples were fragmented by Covaris M220 instrument and size-selected to 300–500 bp using AMPure XP beads. For condition screening, 200 ng fragmented mESC genomic DNA was oxidised by NgTET1 or mTet1CD as describe above and adapter-ligated to Illumina multiplexing adaptor by KAPA HyperPrep kit as described above. Samples were then reduced in a 50 µL reactions with different concentration of pyridine borane and buffer 850 r.p.m. in an Eppendorf ThermoMixer and purified by Zymo-Spin columns. Final libraries were amplified by KAPA Hifi Uracil Plus Polymerase (Kapa Biosystems) and TAPS conversion was tested by TaqαI assay and by next generation screening by paired-end 80 bp sequencing on a NextSeq 500 sequencer (Illumina) using one NextSeq Mid Output kit with 1% PhiX control library spike-in to ~ 10M reads per library. Conditions are summarised in Table 4-7.

Table 4-7. Summary of pyridine borane TAPS conditions tested in first screening sequencing run.

Condition	Oxidation	Pyridine Borane	Buffer	Incubation
PyBr-1	NgTET1	2 M	90 mM NaOAc pH = 5.2	8 h at 50°C
PyBr-2	NgTET1	2 M	90 mM NaOAc pH = 5.2	12 h at 50°C
PyBr-3	NgTET1	2 M	90 mM NaOAc pH = 5.2	16 h at 50°C
PyBr-4	mTet1CD	2 M	90 mM NaOAc pH = 5.2	16 h at 50°C

4.11.19 TAPS optimisation with pyridine borane

500 ng fragmented and size-selected mESC gDNA (300-500 bp) spiked-in with 0.5% CpG methylated lambda DNA were double-oxidised by NgTET1 as described above. For condition screening, 50 ng of oxidised DNA was reduced in a 50 µL reaction containing various amount of pyridine borane compound in different buffers and pH. Reactions were incubated at different temperatures and for different time periods at 850 r.p.m. in an Eppendorf ThermoMixer and purified by Zymo-Spin IC column. TAPS conversion was tested by TaqαI assay on 194mer and/or 216mer control sequences as described above. For some conditions, sequencing libraries were prepared with Kapa HyperPrep kit as described above and TAPS conversion was tested by TaqαI assay before and after final library amplification. Summary of tested conditions is in Table 4-8.

Table 4-8. Summary of optimisation of TAPS conditions with pyridine borane. Sequencing libraries were constructed for conditions in bold (PyBr-15, PyBr-19 and PyBr-21)

Condition	Pyridine Borane	Buffer	Incubation
PyBr-5	1 M	60 mM NaOAc pH = 4.0	16 h at 37 °C
PyBr-6	1 M	60 mM NaOAc pH = 4.3	16 h at 37 °C
PyBr-7	1 M	60 mM NaOAc pH = 4.5	16 h at 37 °C
PyBr-8	2 M	60 mM NaOAc pH = 4.0	16 h at 37 °C
PyBr-9	2 M	60 mM NaOAc pH = 4.3	16 h at 37 °C
PyBr-10	2 M	60 mM NaOAc pH = 4.5	16 h at 37 °C
PyBr-11	1 M	300 mM pH = 4.0	16 h at 37 °C
PyBr-12	1 M	300 mM pH = 4.3	16 h at 37 °C
PyBr-13	1 M	300 mM pH = 4.5	16 h at 37 °C
PyBr-14	1 M	300 mM pH = 4.0	16 h at 37 °C
PyBr-15	1 M	300 mM pH = 4.3	16 h at 37 °C
PyBr-16	1 M	300 mM pH = 4.3	16 h at 25 °C
PyBr-17	1 M	300 mM pH = 4.3	16 h at 32 °C
PyBr-18	1 M	300 mM pH = 4.3	16 h at 40 °C
PyBr-19	1 M	600 mM pH = 4.3	16 h at 37 °C
PyBr-20	1 M	600 mM pH = 4.3	16 h at 40 °C
PyBr-21	0.4 M	600 mM pH = 4.3	16 h at 37 °C
PyBr-22	0.4 M	600 mM pH = 4.3	16 h at 40 °C

4.11.20 TAPS condition with pyridine borane – second screening sequencing

For whole-genome TAPS, mESC genomic DNA was spiked with 0.5% of methylated λ -DNA and 0.025% of unmodified 2 kb spike-in control. DNA samples were fragmented by Covaris M220 instrument and size-selected to 300-500 bp using AMPure XP beads. DNA was additionally spiked 0.25% of N5mCNN and N5hmCNN control oligonucleotides after size-selection with AMPure XP beads. For condition screening, 200 ng fragmented mESC genomic DNA was double-oxidised by NgTET1 and adapter-ligated to Illumina multiplexing adaptor by KAPA HyperPrep kit as described above. Alternatively, 200 ng mESC genomic DNA was first end-repaired and A-tailed and adapter ligated to Illumina multiplexing adapters by KAPA HyperPrep kit and then double-oxidised by mTet1CD as described above. Oxidized DNA (50-100 ng) in 35 μ l of

water was reduced in a 50 µl reaction containing 600 mM sodium acetate solution (pH 4.3) and 1 M pyridine borane for 16 h at 37°C and 850 r.p.m. in an Eppendorf ThermoMixer. The product was purified using Zymo-Spin columns and final libraries were then amplified by KAPA Hifi Uracil Plus Polymerase.

TAPS conversion was tested by TaqαI assay and by next generation screening sequencing and by next generation screening by paired-end 80 bp sequencing on a NextSeq 500 sequencer (Illumina) using one NextSeq Mid Output kit with 1% PhiX control library spike-in. Libraries were sequence to around 10M reads per library.

All conditions are summarised in Table 4-9. Library TAPS-1 that shown the best conversion in screening sequencing and was established as a final TAPS condition was then sequenced deep by paired-end 80 bp sequencing on a NextSeq 500 sequencer (Illumina) using one NextSeq High Output kit with 1% PhiX control library spike-in. Detailed sequencing statistics are shown in Appendix 4 Supplementary Table 4-5.

Table 4-9. Summary of TAPS conditions tested in second screening sequencing run.

Library	Oxidation	Reduction	Buffer	Incubation	PCR cycles
TAPS-1	mTet1CD	1 M PyBr	600 mM pH = 4.3	16 h at 37 °C	5
TAPS-2	NgTET1	1 M PyBr	600 mM pH = 4.3	16 h at 37 °C	5

4.11.21 WGBS

For WGBS, mESC genomic DNA was spiked-in with 0.5% of unmethylated λ-DNA were used. DNA samples were then fragmented by Covaris M220 instrument and size-selected to 200–400 bp using AMPure XP beads. WGBS library was prepared from 200 ng of fragmented mESC genomic DNA. End-repair and A-tailing reaction and ligation of methylated adapter (NextFlex) were prepared with KAPA HyperPlus kit (Kapa Biosystems) according to manufacturer’s protocol. Subsequently, DNA underwent bisulfite conversion with EpiTect Bisulfite Kit (Qiagen) according to Illumina’s protocol. The final library was amplified with KAPA Hifi Uracil Plus Polymerase (Kapa

Biosystems) for 6 cycles and cleaned up on 1× AMPure XP beads. The WGBS sequencing library was paired-end 80 bp sequenced on a NextSeq 500 sequencer (Illumina) using one NextSeq High Output kit with 15% PhiX control library spike-in. Detailed sequencing statistics are shown in Appendix 4 Supplementary Table 4-5.

4.11.22 DNA damage test on fragments with different length

mESC genomic DNA was spiked with 0.5% of CpG-methylated λ -DNA and left unfragmented or sonicated with a Covaris M220 instrument and size-selected to 500-1 kb or 1-3 kb using AMPure XP beads. DNA (200 ng) was single oxidized with mTet1CD and reduced with pyridine borane complex as described above or converted with EpiTect Bisulfite Kit (Qiagen) according to manufacturer's protocol. DNA (10 ng) before and after TAPS and bisulfite conversion was run on a 1% agarose gel. To visualize bisulfite conversion, gel was cooled down for 10 min in an ice bath. 5mC conversion in TAPS samples was tested by Taq α I digestion assay

4.11.23 WGBS data processing

Paired-end reads were downloaded as FASTQ files from Illumina BaseSpace and subsequently quality trimmed with Trim Galore! v0.6.0 (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/). Read pairs where at least one read was shorter than 35 bp after trimming were removed. Trimmed reads were mapped to a genome combining the mm9 version of the mouse genome, λ -phage and PhiX (sequence from Illumina iGENOMES) using Bismark v0.22.3 and the --no_overlap option (Krueger and Andrews, 2011). The 'three-C' filter was used to remove reads with excessive non-conversion rates. PCR duplicates were called using Picard v2.20.2 MarkDuplicates (<http://broadinstitute.github.io/picard/>). Regions known to be prone to mapping artefacts were downloaded

(<https://sites.google.com/site/anshulkundaje/projects/blacklists>) and excluded from further analysis (Dunham *et al.*, 2012).

4.11.24 TAPS data pre-processing

Paired-end reads were downloaded from Illumina BaseSpace and subsequently quality trimmed with Trim Galore! v0.6.0. Read pairs where at least one read was shorter than 35 bp after trimming were removed. Trimmed reads were mapped to a genome combining spike-in sequences, and the mm9 version of the mouse genome using asTair v3.2.4 (<https://bitbucket.org/bsblabludwig/astair/>). PCR duplicates were called using Picard v2.20.2 MarkDuplicates (<http://broadinstitute.github.io/picard/>). Regions known to be prone to mapping artefacts were downloaded (<https://sites.google.com/site/anshulkundaje/projects/blacklists>) and excluded from further analysis (Dunham *et al.*, 2012). Modified bases were detected using deduplicated and uniquely mapped reads using asTair v3.2.4 (<https://bitbucket.org/bsblabludwig/astair/>).

4.11.25 Sequencing quality analysis of TAPS and WGBS

Base content per read position and quality score statistics per nucleotide were analysed by FastQC v0.11.7 from trimmed FASTQ files and visualised in R.

4.11.26 Coverage analysis of TAPS and WGBS

Per-base genome coverage files were generated with Bedtools v2.25 genomecov (Quinlan and Hall, 2010). To compare the relative coverage distributions between TAPS and WGBS, TAPS reads were subsampled to the corresponding coverage median in WGBS using the `-s` option of samtools view.

4.11.27 Analysis of cytosine modifications measured by TAPS and WGBS

The fraction of modified reads per base for WGBS and TAPS was calculated using Bismark and asTair, respectively. Genetic variants at CpG sites specific to the E14 cell line were excluded from the analysis (Incarnato, Krepelova and Neri, 2014). Bedtools intersect, bedtools map, statistical analyses and figures were generated in python3 and R. To plot the modification levels around CGIs, all CGI coordinates for mm9 were downloaded from the UCSC genome browser, binned into 20 windows, and extended by up to 50 windows of size 80 bp on both sides (as long as they did not reach half the distance to the next CGI). Average modification levels (in CpGs) and coverage (in all bases, both strands) in each bin were computed using bedtools map. The values for each bin were again averaged and subsequently plotted in R.

*The good-news stories in medicine are
detection, early intervention.*

Thomas R. Insel

5 Methylation profiling in cell-free DNA by TAPS – proof-of-concept study

5.1 Introduction

Early cancer diagnosis is crucial for improvement of patient prognosis and survival (Greten *et al.*, 2005; McPhail *et al.*, 2015; Hur *et al.*, 2016; Stark *et al.*, 2016; Shah *et al.*, 2017). Unfortunately, early detection methods are not available for most cancer types and many patients are still diagnosed at an advanced, symptomatic stage (McPhail *et al.*, 2015; Hur *et al.*, 2016). Pancreatic ductal adenocarcinoma (PDAC) and hepatocellular carcinoma (HCC) are two of the deadliest cancer types for which early detection diagnostic methods are yet to be developed (Greten *et al.*, 2005; Stark *et al.*, 2016; Shah *et al.*, 2017; Neoptolemos *et al.*, 2018; Rawla, Sunkara and Gaduputi, 2019). Both PDAC and HCC are routinely diagnosed by imaging techniques followed by tissue biopsy but these approaches are not suitable for early screening (Pocha *et al.*, 2013; Hur *et al.*, 2016; Ngamruengphong and Canto, 2016). Although blood-based protein biomarkers, CA19-9 and AFP, are clinically used to facilitate diagnosis and monitoring of PDAC and HCC, they have limited sensitivity and specificity. Therefore, new improved diagnostic assays for PDAC and HCC are still needed.

Liquid biopsy is considered a promising approach for development of novel, minimally invasive methods for cancer detection. In particular, analysis of genetic and epigenetic changes in cell-free DNA (cfDNA) could facilitate early cancer diagnosis (Wan *et al.*,

2017). On the other hand, most of proposed cfDNA-based diagnostic assays have not achieved sensitivity and specificity required for clinical application (Watts, 2018).

Recent studies suggest that methylation biomarkers can detect cancer with high sensitivity and specificity as well as provide information for accurate prediction of tissue of origin and help to identify tumour location (Shen *et al.*, 2018; Nassiri *et al.*, 2020; Nuzzo *et al.*, 2020, Sun *et al.*, 2015; Guo *et al.*, 2017; Moss *et al.*, 2018). In addition, other cfDNA features including fragmentation profiles and genetic changes are affected in the presence of malignancy and aberrant fragmentation patterns can facilitate sensitive cancer prediction (Jiang *et al.*, 2015; Cohen *et al.*, 2018; Mouliere *et al.*, 2018; Cristiano *et al.*, 2019). Comprehensive epigenetic and genetic analysis of cfDNA in healthy and disease could help to develop novel, sensitive approaches for cancer detection. Unfortunately, simultaneous genome-wide epigenetic and genetic analysis with current gold standard – bisulfite sequencing remains challenging (Tanaka and Okamoto, 2007; Worm Ørntoft *et al.*, 2017).

5.2 Motivation

Despite promising performance of methylation-based biomarkers in exploratory studies, most of them have not been translated into clinical use (Koch *et al.*, 2018). Lack of sensitive and specific methylation biomarkers ready for bench-to-bed transfer can be at least partially caused by technological challenges in genome-wide methylation analysis in cfDNA. Most of the current biomarkers were identified based on tumour data and do not include potential cancer-specific methylation changes in cfDNA derived e.g. from immune cells (Koestler *et al.*, 2012; Y. Zhang *et al.*, 2018). Identification of such biomarkers would require comprehensive genome-wide cfDNA methylation sequencing which is difficult with the gold standard bisulfite sequencing due to an extensive DNA

damage and significant loss of input samples introduced by bisulfite treatment as well as low level of cfDNA in circulation (Worm Ørntoft *et al.*, 2017). In addition, current methylation sequencing methods do not provide information about other cfDNA characteristics that could further improve accuracy of cancer detection. For instance, DNA damage introduced by bisulfite treatment can interfere with natural cfDNA fragmentation profiles and affect simultaneous methylation and fragmentation analysis (Tanaka and Okamoto, 2007; Worm Ørntoft *et al.*, 2017; Erger *et al.*, 2020).

Recently, we developed a new, bisulfite-free method to analyse methylome called TAPS. Owing to its mild character we sought to optimise this approach for cfDNA analysis and applied it to the clinical cfDNA samples from PDAC, HCC and non-cancer controls for multidimensional, genome-wide characterization of cfDNA methylation and fragmentation profiles.

5.3 TAPS provides reliable, base-resolution cfDNA methylation landscape

To enable bisulfite-free, base resolution methylation sequencing in cfDNA we had to optimise TAPS protocol to work with low input material. We could generate TAPS libraries from as low as 1 ng of genomic DNA and cfDNA with minimal number of PCR cycles using unmodified TAPS protocol (Figure 5-1, a-b). Moreover, adding carrier DNA to ligated cfDNA TAPS libraries prior conversion helped to improve final library yield, possibly due to improved sample recovery during clean-up steps (Figure 5-1, c). Thus, this approach was chosen as the final cfDNA TAPS protocol.

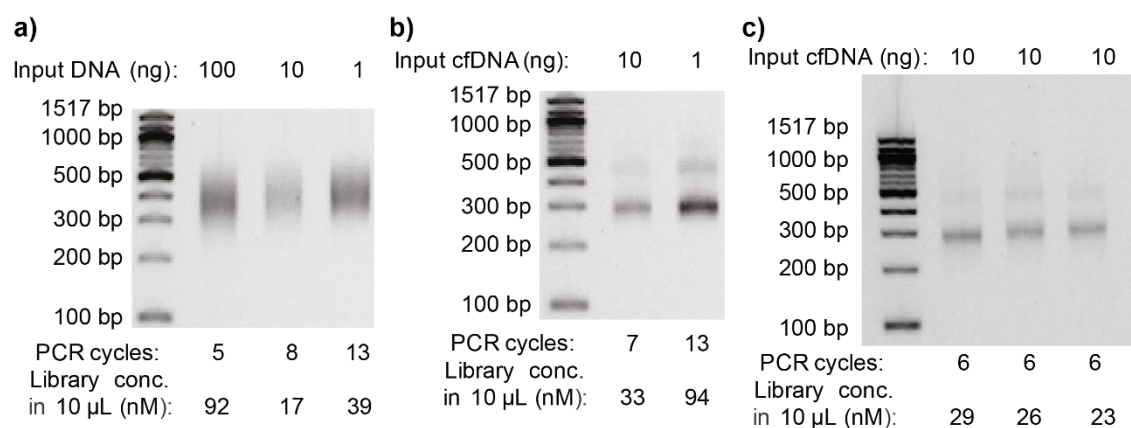


Figure 5-1. Representative images of low input TAPS libraries prepared with dsDNA KAPA HyperPrep library preparation kit. (a) mESC gDNA TAPS libraries. (b) cfDNA TAPS libraries. (c) cfDNA TAPS libraries. Conversion performed in the presence of carrier DNA.

To assess TAPS performance on cfDNA, we sequenced three libraries prepared according to the optimised protocol. We confirmed that cfDNA TAPS libraries were at high quality with high 5mC conversion ($90.74\% \pm 0.35\%$) and low false positive rate ($0.17\% \pm 0.03\%$) (Figure 5-2, a). Lower 5mC conversion observed in this pilot sequencing compared to other TAPS samples was probably caused by a lower efficiency of mTet1CD oxidation due to the long-term storage of the enzyme batch used for preparation of these samples.

CfDNA TAPS libraries exhibited good mapping efficiency and low duplication rate. At ~ 10 X sequencing, 91.15% (SD = 1.01%) of total reads were uniquely mapped and 85.49% (SD = 1.74%) remained after removal of PCR duplicates, which indicates that our method generated sequencing libraries at high complexity even from as little as 10 ng of cfDNA (Figure 5-2, b). Detailed mapping statistics are summarised in Appendix 5 Supplementary Table 5-1.

This together demonstrated that TAPS method could be easily adopted for genome-wide methylation sequencing in cfDNA with overall performance similar to this observed in high input genomic samples.

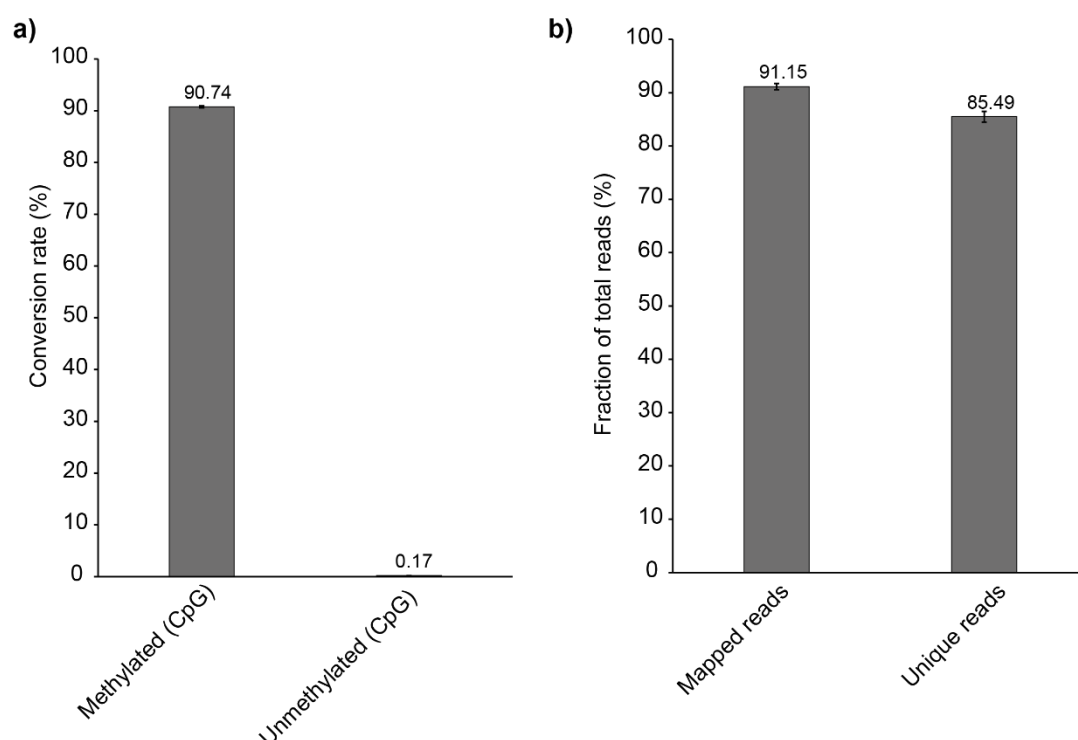


Figure 5-2. Conversion and mapping assessment of three pilot cfDNA TAPS libraries prepared with carrier DNA protocol. (a) cfDNA TAPS libraries conversion of 5mC based on CpG-methylated lambda spike-in (Methylated CpG) and unmodified cytosine based on unmodified 2 kb spike-in (unmethylated CpG). Error bars represent standard error. Positions covered by less than 3 reads were excluded from analysis. (b) Unique mapping rate and unique deduplicated mapping rate in the pilot cfDNA TAPS libraries. Error bars represent standard error.

Having confirmed that TAPS is compatible with cfDNA sequencing, we applied this protocol to 96 clinical samples in a proof-of concept study to investigate ability of our method to provide multidimensional information about cfDNA methylation and fragmentation profiles in cancer and control subjects and to test whether cfDNA characteristics retrieved from TAPS sequencing could have a predictive value. Our study cohort included 22 HCC, 31 PDAC and 30 non-cancer controls (referred here as healthy) together with 9 pancreatitis and 4 liver cirrhosis patients. Pancreatitis and liver cirrhosis are chronic conditions that are associated with an increased risk of development of PDAC and HCC, respectively (Neoptolemos *et al.*, 2018; Lim and Singal, 2019) and they can serve as an additional validation set for our prediction models.

CfDNA TAPS libraries were prepared in 4 separate experimental batches. All batches were designed to have similar distribution of age, gender, and disease status (Figure 5-3, a-c). CfDNA TAPS libraries were prepared from 10 ng of DNA and amplified for 7 PCR cycles. All 96 samples were constructed into successful libraries at expected size and with average yield 24 nM (SD = 7.97 nM) (Figure 5-3, d-e). CfDNA TAPS libraries were quantified by qPCR, pooled together and sequenced on NovaSeq 6000. In average we obtained 363368387 (SD = 62880258) total reads and 278223223 (SD = 44933850) uniquely mapped, deduplicated reads, which corresponds to the average sequencing depth ~ 11.7X. Detailed mapping statistics are shown in Appendix 5 Supplementary Table 5-2. There was 2.8-fold difference between the most and the least covered sample and 2-fold difference after removal of this high coverage outlier (Figure 5-3, f). Significantly high coverage of this sample was likely caused by a pipetting error during preparation of sequencing pool.

We then performed quality assessment of prepared TAPS libraries. We specifically looked at 5mC conversion and false positive rate, genomic coverage and batch effect to exclude possibility of technical issues that could affect downstream analysis.

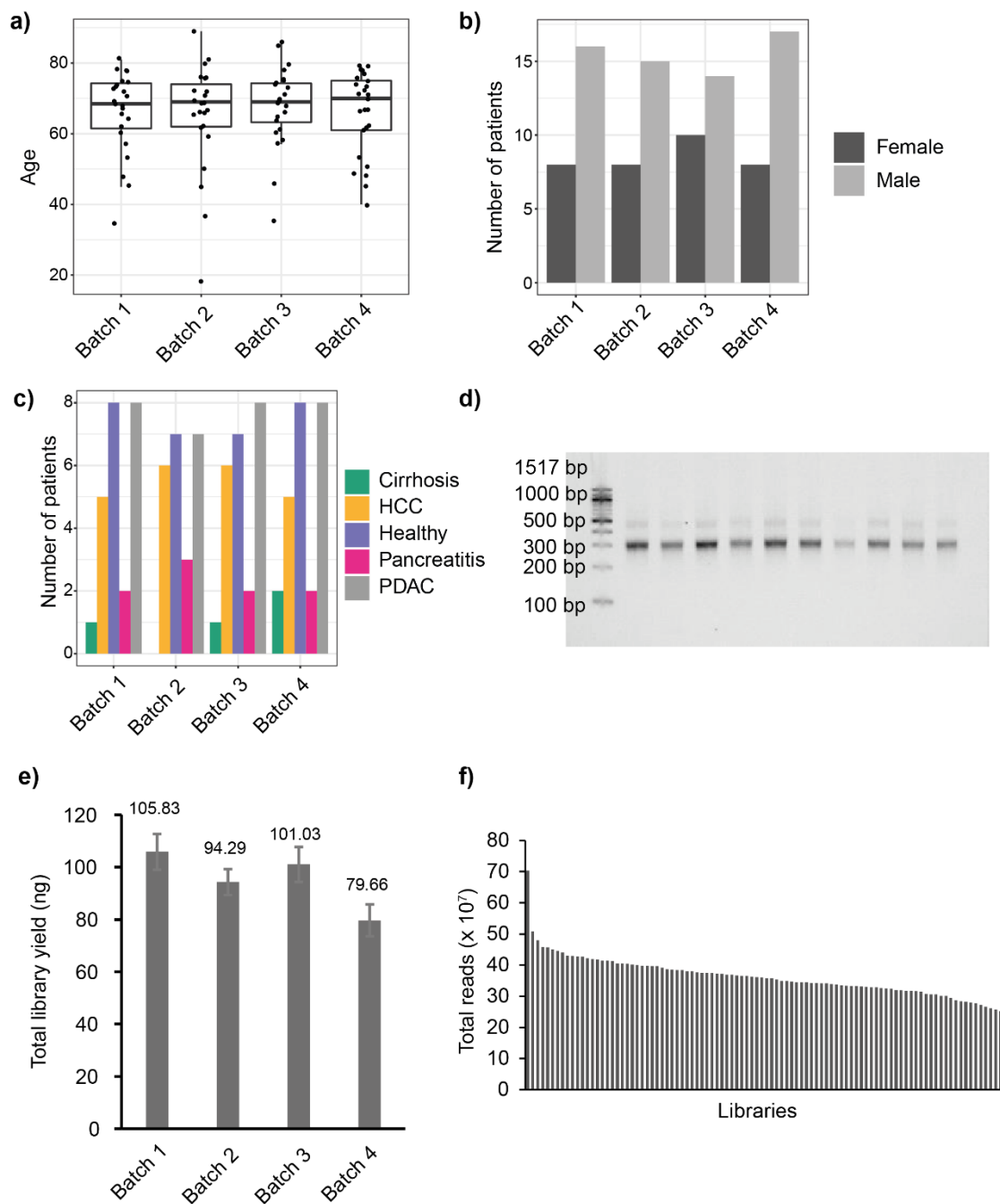


Figure 5-3. (a) Boxplot of patient age distribution in 4 sample batches. (b) Bar chart of number of female and male samples processed in each batch. (c) Number of samples of each disease state processed in each batch. (d) Representative 2% agarose gel of cfDNA TAPS libraries. (e) Comparison of average total library yield for each batch. Error bars represent standard error. (f) Bar chart of total read number per library for 96 cfDNA samples.

Average conversion of 5mC at CpG context based on the CpG methylated spike-in control was 96.7% (SD = 2.43) and false positive rate at CpG context was 0.28% (SD = 0.02). There were not significant differences in the conversion rate depending on sample batch, subject age, gender or disease status (Figure 5-4, a-d). Two samples exhibited low

5mC conversion below 85% and were excluded from analysis. Since these samples were processed in two different batches, we attributed low conversion to the random technical error during TET oxidation or pyridine borane conversion.

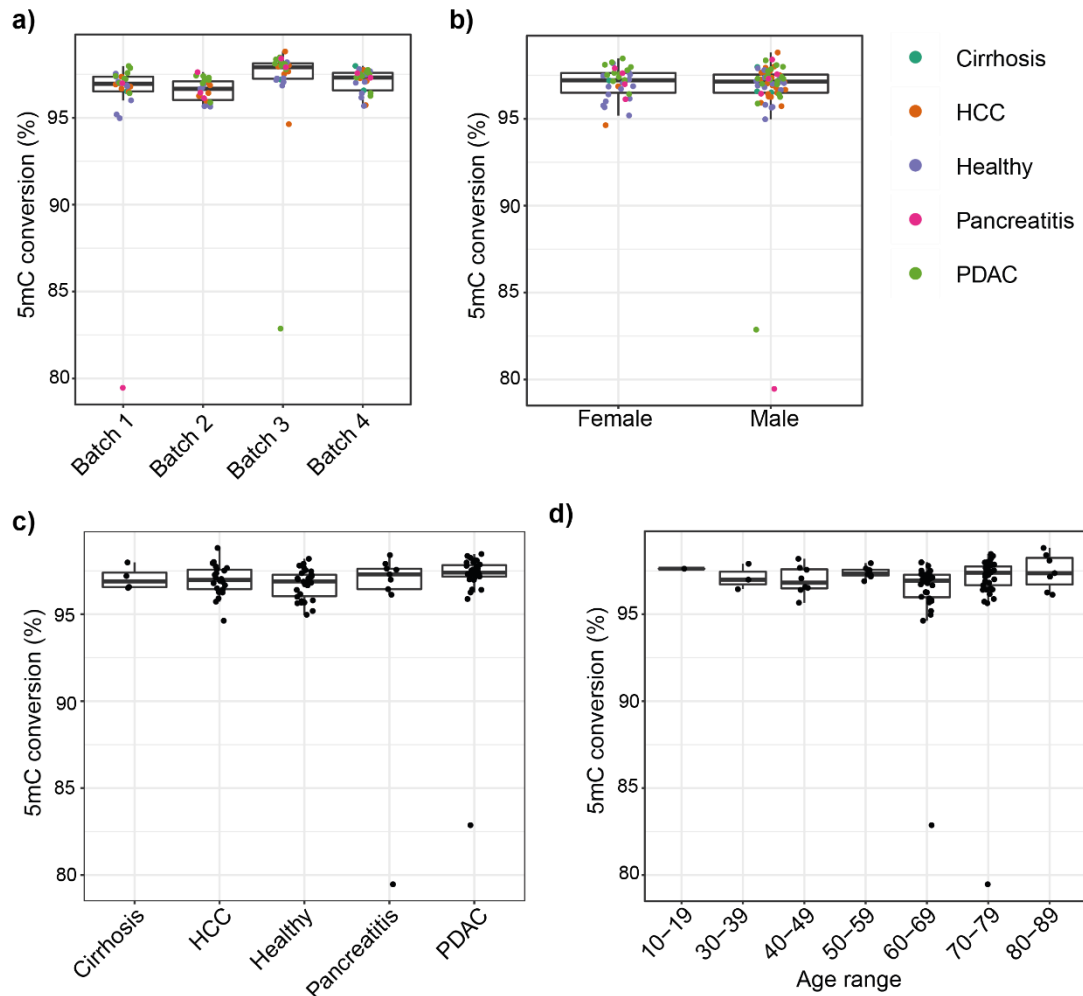


Figure 5-4. Boxplots of 5mC conversion at CpG context depending on (a) sample batch, (b) gender, (c) disease status and (d) age range. 5mC conversion was calculated based on CpG-methylated lambda spike-in control.

We then sought to find whether batch, age, conversion or gender caused any bias in the data at the genome-wide level. To investigate it, we calculated weighted methylation in 1 kb genomic windows and perform unsupervised analysis by PCA. Analysed samples did not separate based on sample batch, age or gender indicating lack of a significant source of bias due to cohort design or 5mC conversion (Figure 5-5, a-c). On the other hand, we found single pancreatitis sample that was one of the outliers in PCA plot and

clustered separately from other pancreatitis samples (Figure 5-5, b-d). This particular sample was collected from an 18 years old individual and was significantly younger than the rest of the cohort (Figure 5-6). Mean age of all patients involved in our study was 66.24 years (SD = 12.48). The age of this patient was more than 3 standard deviations outside of mean age in the cohort thus could be defined as an age outlier. Since methylation signatures are age-dependent (Xiao, Wang and Kong, 2019) this age difference could affect downstream methylation analysis and therefore we decided to exclude this sample from further analysis.

Moreover, after revising the clinical data of HCC patients we discovered that one of the patients underwent liver transplant before recruitment to the study and did not have an active HCC at the time of sample collection. For this reason, we excluded this patient from downstream analysis.

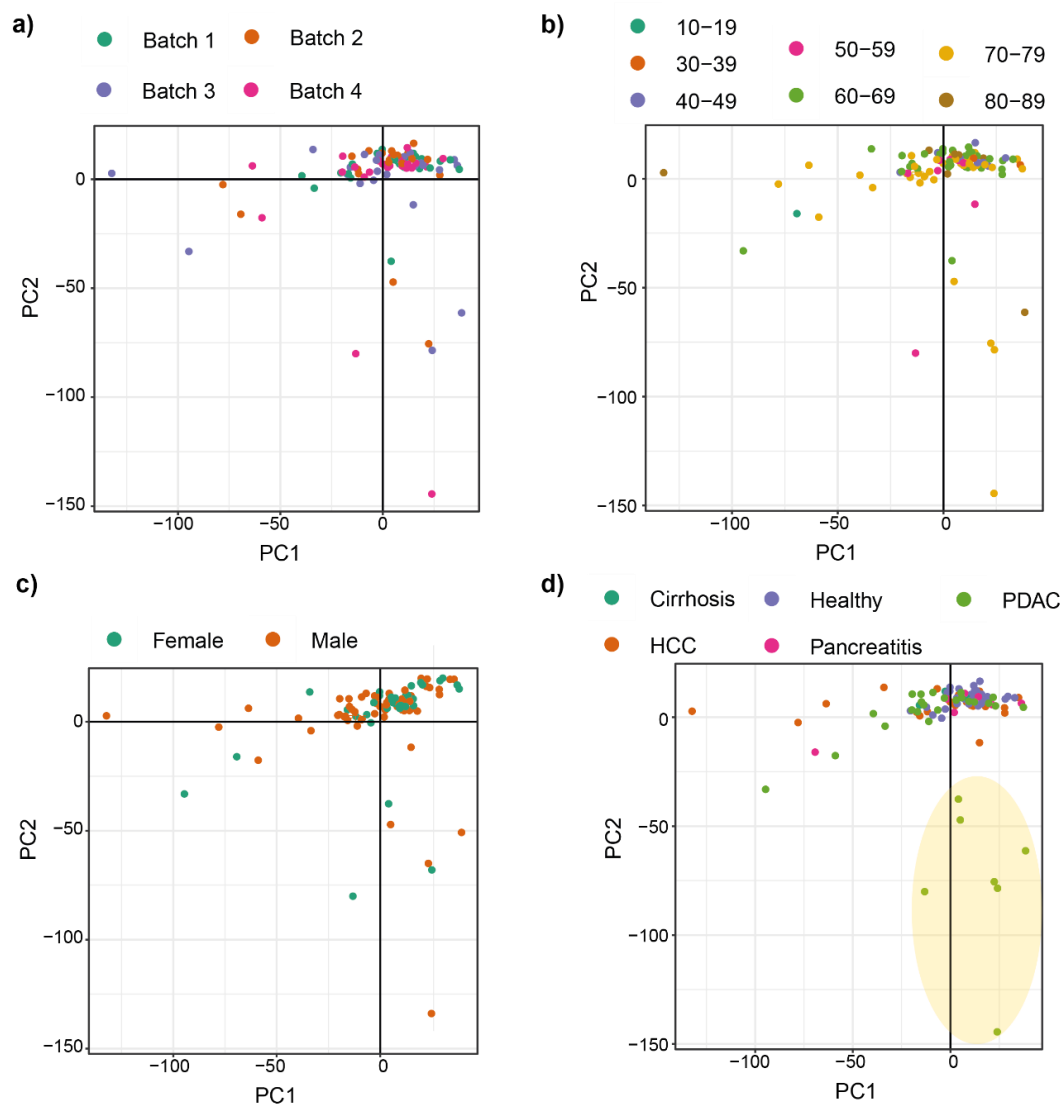


Figure 5-5. PCA analysis of cfDNA methylation in 1000 bp windows. Samples are coloured by (a) batch, (b) age, (c) gender or (d) status. In figure (d) PDAC outliers are marked by yellow circle.

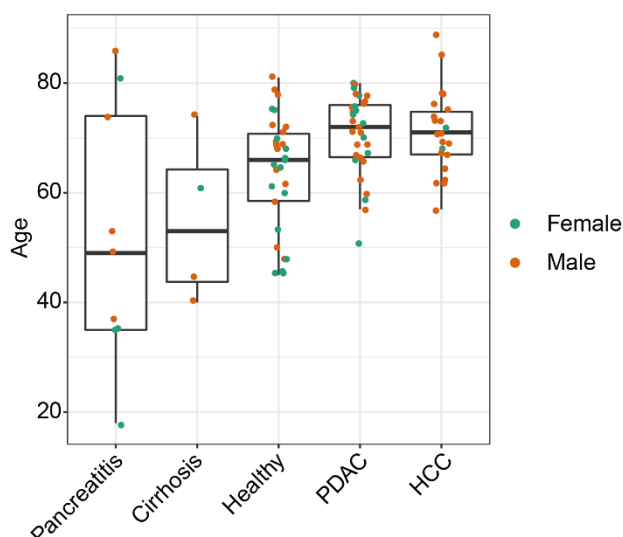


Figure 5-6. Boxplot of age and gender distribution of 96 patients initially included in cfDNA TAPS cohort.

In addition, we found that seven samples collected from pancreatic cancer patients clustered together separately from other samples (Figure 5-5, d). We found that these seven samples exhibited distinct per CpG coverage distribution compared to the remaining samples (Figure 5-7, a-d). Specifically, PDAC outlier samples had broad per CpG coverage distribution with a high proportion of non-covered and highly covered (coverage ≥ 50) CpG sites. When viewed in Genome Browser these samples exhibited uneven coverage with dips corresponding to the non-covered regions (Figure 5-7, e). To note, average sequencing depth for these samples did not differ from other samples thus differences in the total read number cannot explain observed issues with coverage distribution (Figure 5-7, f). Additionally, we found that PDAC outlier libraries had an increased GC content compared to other libraries (Figure 5-7, g). Average GC content of sequenced libraries was 41.14% (SD = 2.20%). GC content in 6 outlier samples was more than 3 standard deviations from mean and in one sample more than 2 standard deviations from mean. Therefore, these samples can be defined as GC content outliers.

We did not find any clinical or biological details that could explain these observations, however, findings of initial sequencing quality control suggest potential issue with quality and integrity of cfDNA. Importantly, all affected samples had been consecutively collected over a period of two months by one nurse and our clinical collaborator indicated that possible problems during blood collection, blood storage or plasma processing cannot be excluded. Since these 7 libraries exhibited distinct, unexplained coverage distribution and GC content clearly different from other samples in the cohort and we were not able to exclude potential issue with cfDNA quality and cfDNA degradation as a source of these observations, we decided to exclude them from further analysis.

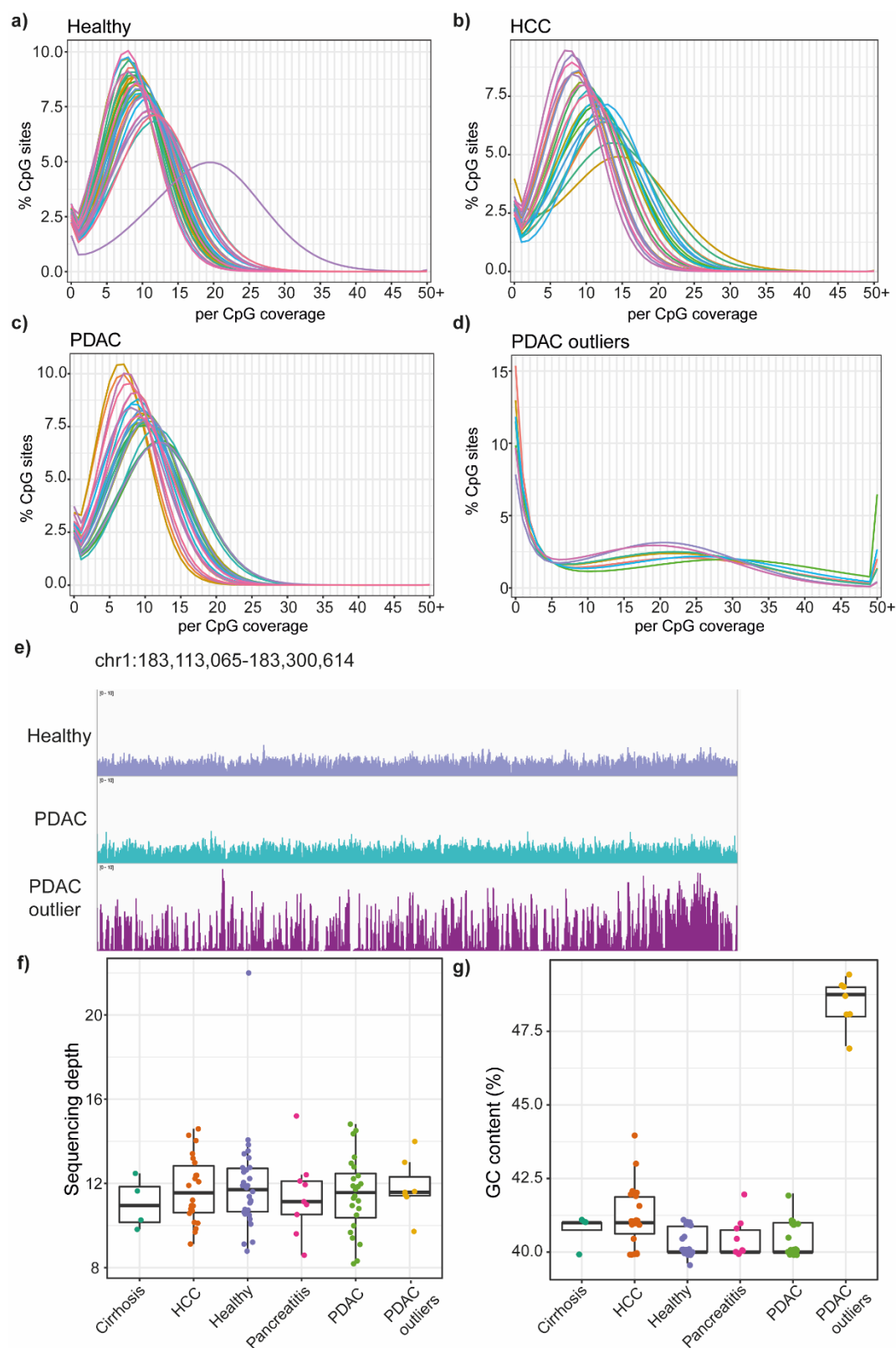


Figure 5-7. Per CpG coverage distribution in cfDNA TAPS libraries for (a) Healthy (b) HCC, (c) PDAC, (d) PDAC outliers. Each line represents individual sample. Number of CpG sites covered by 50 or more reads were summed as one bin 50. One healthy library that does not align with other samples and has broader coverage distribution skewed towards higher coverage had twice as high number of total reads compared to other libraries due to pooling error. (e) IGV view

of coverage of representative healthy, PDAC and PDAC outlier sample. (f) Boxplot of sequencing depth distribution of cfDNA TAPS libraries. (g) Boxplot of GC content of cfDNA TAPS libraries.

After initial quality assessment 85 out of initial 96 samples were included to downstream analysis. Clinical characteristics of the final cohort after quality control are summarised in Table 5-1. Detailed clinical characteristics are listed in Appendix 5 Supplementary Tables 5-3 – 5-7. Gender and age distribution in the final cohort is shown in Figure 5-8.

Table 5-1. Clinical characteristic of samples.

	HCC (n=21)	PDAC (n=23)	Pancreatitis (n=7)	Cirrhosis (n=4)	Healthy (n=30)
General characteristics					
Female	1 (4.76%)	9 (39.13%)	3 (42.85%)	1 (25%)	15 (50%)
Male	20 (95.24 %)	14 (60.87%)	4(57.14%)	3 (75%)	15 (50%)
Age	71.14 ± 7.76 (57-89)	70.70 ± 6.82 (57-80)	53.71± 21.56 (35-86)	55 ± 15.51 (40-74)	63.93 ± 10.54 (45-81)
Tumour size (mm)	49.67 ± 42.22 (9-160)	35.87 ± 11.22 (12-60)	NA	NA	NA
TNM classification					
T1N0M0	7 (33.33%)	1 (4.35%)	NA	NA	NA
T1cN0M0	0	1 (4.35%)	NA	NA	NA
T1NxM0	1 (4.76%)	0	NA	NA	NA
T2N0M0	4 (19.05%)	1 (4.35%)	NA	NA	NA
T2N1M0	0	5 (21.74%)	NA	NA	NA
T2N2M0	0	6 (26.09%)	NA	NA	NA
T2NxM0	2 (9.53%)	0	NA	NA	NA
T3aN0M0	4 (19.05%)	0	NA	NA	NA
T3bN0M0	1 (4.76%)	0	NA	NA	NA
T3N1M0	0	3 (13.04%)	NA	NA	NA
T3N2M0	0	4 (17.39%)	NA	NA	NA
T3aNxM0	1 (4.76%)	0	NA	NA	NA
T3N1M1	0	1 (4.35%)	NA	NA	NA
T3N2Mx	0	1 (4.35%)	NA	NA	NA
T4NxM1	1 (4.76%)	0	NA	NA	NA
Stage					
I	8 (38.10%)	3 (13.04%)	NA	NA	NA
II	6 (28.57%)	8 (34.78%)	NA	NA	NA
III	6 (28.57%)	11 (47.83%)	NA	NA	NA
IV	1 (4.76%)	1 (4.35%)	NA	NA	NA

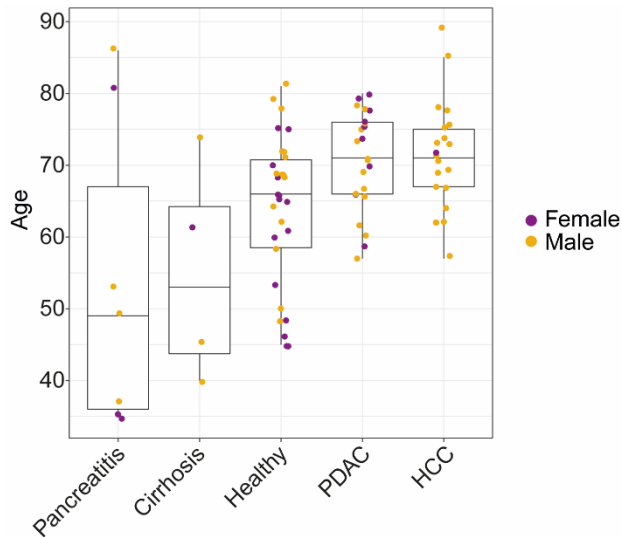


Figure 5-8. Boxplot of age and gender distribution of 85 patients from cfDNA TAPS cohort included into downstream analysis.

CfDNA concentrations were reported to be higher in cancer patients compared to healthy and in more advanced disease stage (Wan *et al.*, 2017). We sought to check whether we could see similar trend in our cohort. We calculated cfDNA yield in plasma as the amount of cfDNA in nanograms (ng) isolated from 1 mL of plasma. We found significant increase in plasma cfDNA yield in the samples collected from HCC and PDAC patients compared to the healthy controls (HCC vs. healthy - P value: 0.0001; PDAC vs. healthy - P value: 0.04805; Figure 5-9, a). Moreover, we found weak, however, statistically non-significant correlation between both tumour stage and tumour size and cfDNA yield in HCC and to a lesser extent between tumour stage and cfDNA yield in PDAC (Figure 5-9, b-e). Increased cfDNA concentration in plasma of cancer patients, especially in more advanced disease, may be associated with higher tumour contribution to cfDNA or tissue damage and remodelling in the affected organs.

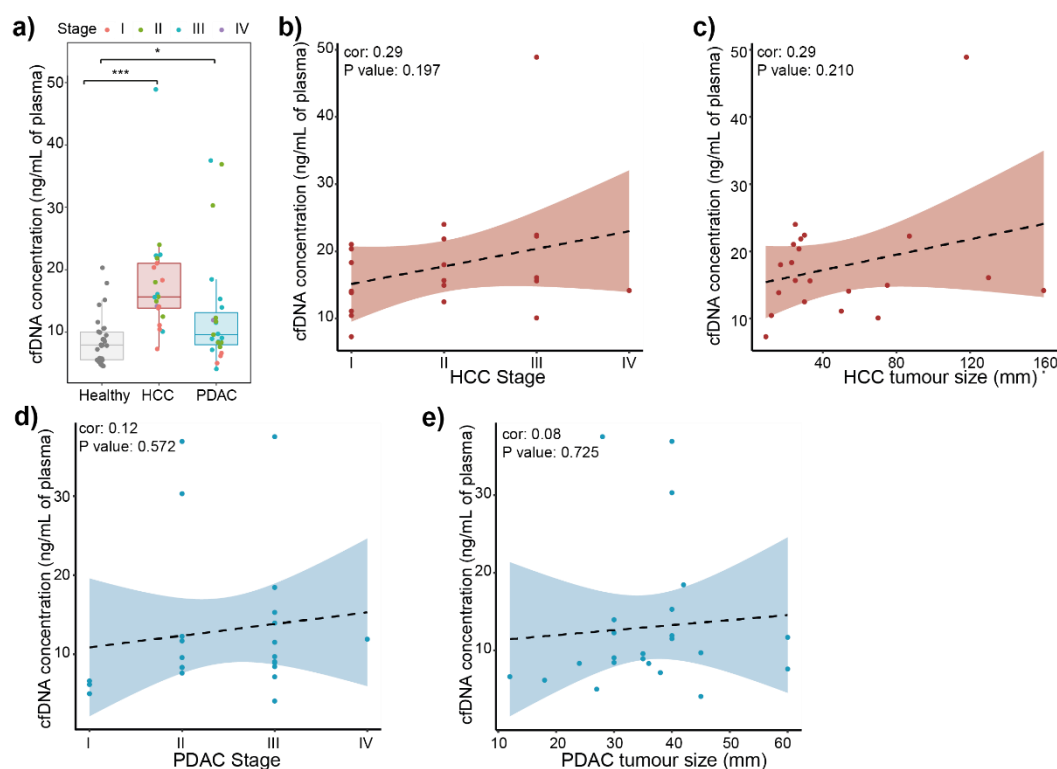


Figure 5-9. Comparison of plasma cfDNA concentrations between cancer patients and healthy controls. (a) Boxplot of cfDNA yield defined as amount of cfDNA isolated from 1 mL of plasma between cancer and healthy. cfDNA yield in healthy was significantly lower ($8.7 \text{ ng/mL} \pm 3.96 \text{ ng}$) than in PDAC ($13 \text{ ng/mL} \pm 9.35 \text{ ng}$; P value: 0.048) and HCC ($17.75 \text{ ng/mL} \pm 8.49 \text{ ng}$; P value: 0.0001). Statistical significance was assessed with two-tailed t-test. (b-c), Correlation plots of cfDNA concentration in plasma in HCC patients and (b) tumour stage and (c) tumour size (mm). (d-e), Correlation plots of cfDNA concentration in plasma in PDAC patients and (d) tumour stage and (e) tumour size. Each dot represents an individual patient. Dashed lines represent the linear trends fitted with linear regression. Shaded areas represent 95% confidence intervals of the fitted models. Pearson correlation coefficients (cor) and P values are shown in the plots.

We started cfDNA analysis from general characterization of methylation profiles in PDAC, HCC and non-cancer controls. cfDNA CpG methylation showed bimodal distribution with most CpG sites in cfDNA fully methylated and highly methylated (methylation value > 0.75) or unmethylated. No significant differences in the global methylation distribution were observed between cancer and non-cancer patients (Figure 5-10, a-c). On the other hand, few cancer samples, especially HCC, exhibited decreased methylation level across chromosomes which may reflect global hypomethylation characteristic for cancer and reported in cfDNA from cancer patients in previous studies (Figure 5-10, d-f) (Allen Chan *et al.*, 2013). We also found higher level of methylation

variation in HCC and PDAC group compared to healthy, which suggest high methylation heterogeneity in cancer patients (Figure 5-10, g).

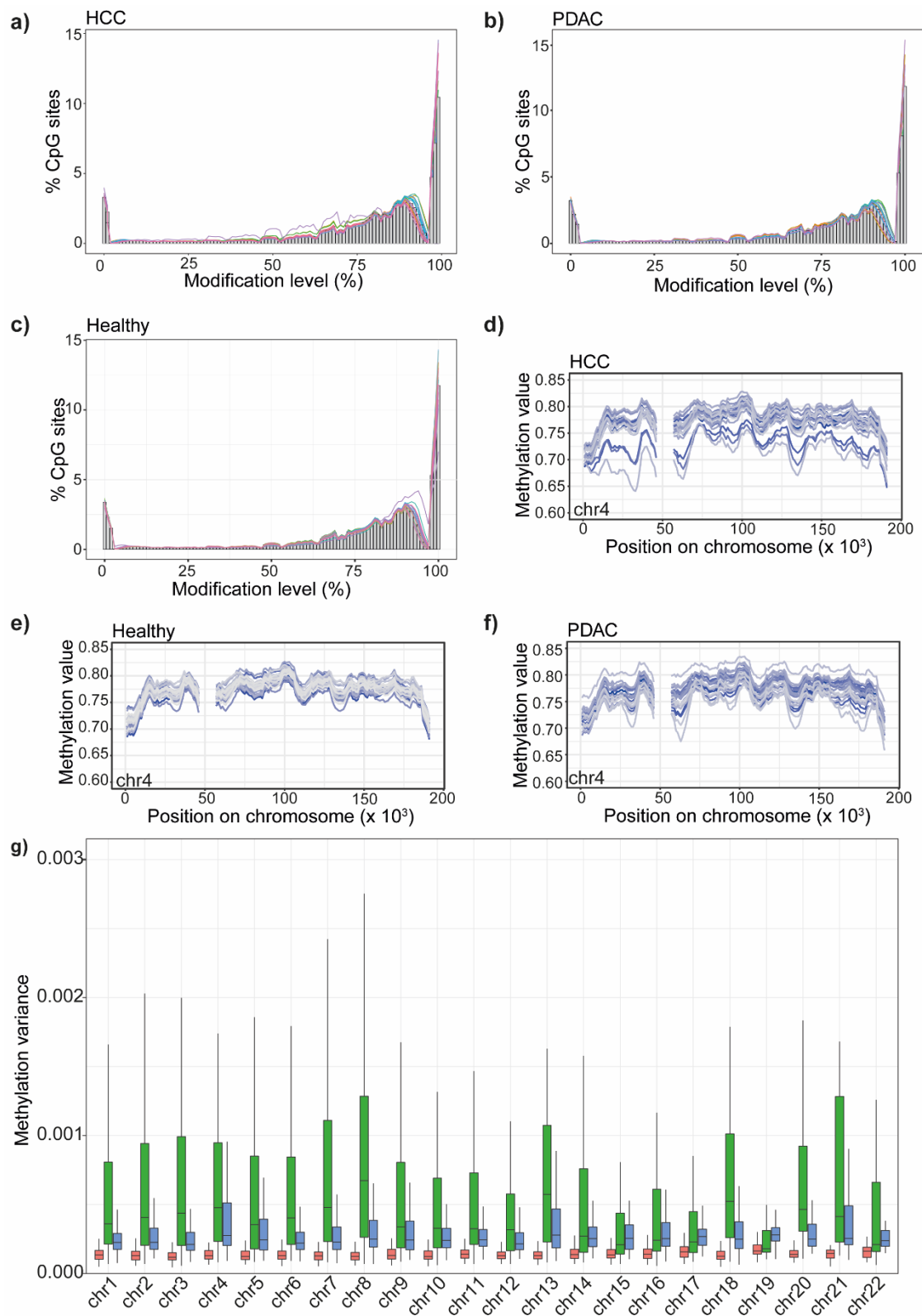


Figure 5-10. Bar charts of genome-wide cfDNA average CpG modification distribution in (a) HCC, (b) PDAC and (c) healthy controls. Overlaid line plots represent average CpG methylation

distribution for individual sample. CpG sites covered by less than 3 reads were excluded from analysis. CpG methylation was Gaussian-smoothed (smoothing window size 5). Distribution of modification levels in chromosome 4 for in (d) HCC, (e) Healthy and (f) PDAC. Each line represents individual sample. Average CpG modification value was calculated per 100 kb windows along chromosome 4 and Gaussian-smoothed (smoothing window size 10). (g) Boxplot of 1 Mb bins methylation variance in healthy and disease.

We then performed unsupervised analysis of cfDNA methylation in various genomic features and we found that methylation signatures in regulatory elements including promoters and enhancers showed promising separation between cancer and control samples with majority of cancer and healthy samples displaying different distributions on PCA plots (Figure 5-11, a-d).

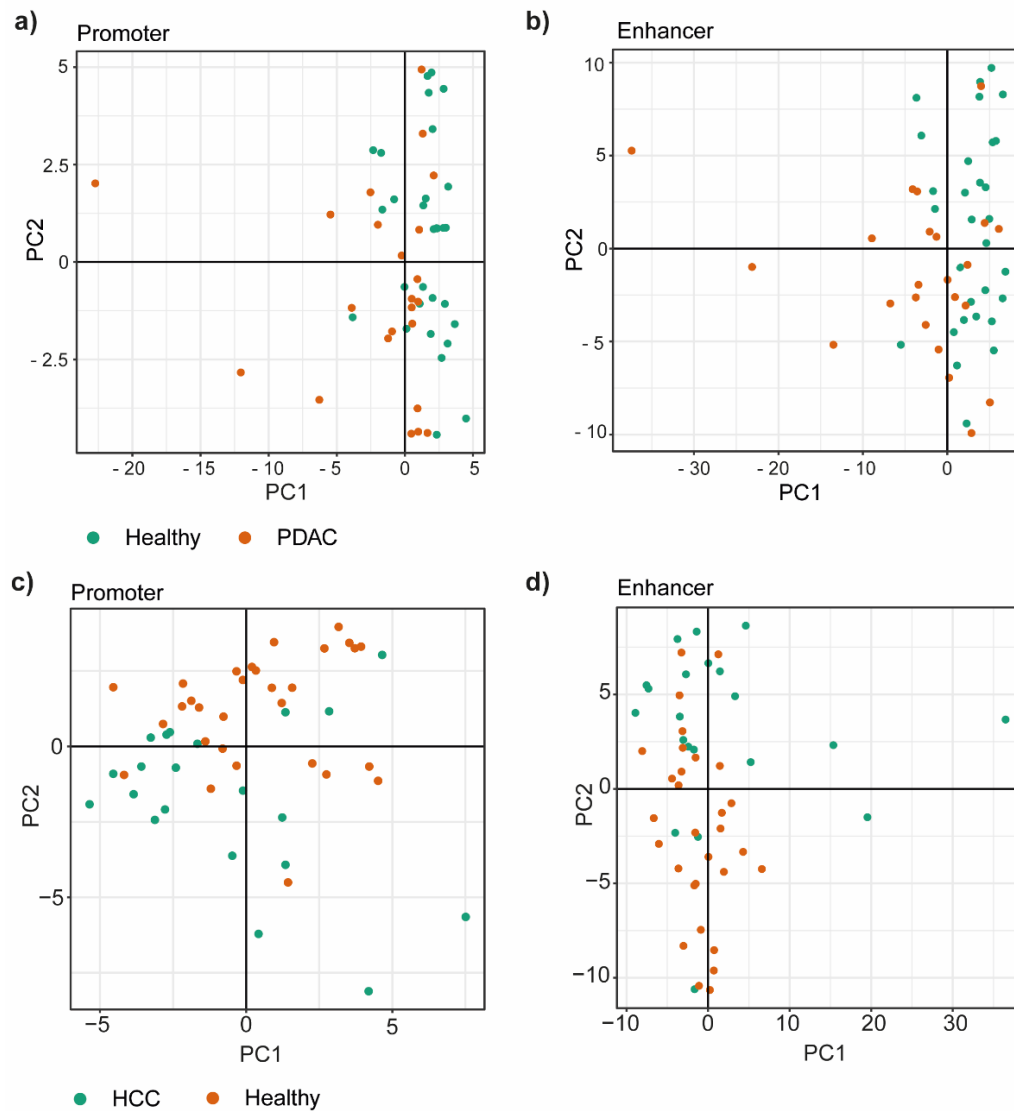


Figure 5-11. PCA analysis of cfDNA promoter methylation (a) and enhancer methylation (b) in PDAC and healthy. PCA analysis of cfDNA promoter methylation (c) and enhancer methylation (d) in HCC and healthy.

Subsequently, we looked at the differentially methylated CpG sites in PDAC, HCC and non-cancer controls in a genome-wide, supervised manner.

We found 59569 differentially methylated CpG sites (DMCs) between PDAC and control group (P value ≤ 0.01 , methylation difference > 0.1), including 38351 hypermethylated sites and 21218 hypomethylated sites (Figure 5-12, a). Differentially methylated sites were enriched in CGI shores and shelves, protein coding regions as well as in regulatory elements including promoters, enhancers and TF binding sites (Figure 5-12, b). Gene ontology analysis revealed that differentially methylated sites are associated with genes involved in function of blood cells (e.g. haemostasis, platelets activation, Fc gamma receptor-mediated phagocytosis) as well as important signalling pathways including EGFR receptor and c-Kit receptor signalling pathways with known involvement in pancreatic cancer (Yasuda *et al.*, 2006; Navas *et al.*, 2012) (Figure 5-12, c). This suggests that differential methylation signal in pancreatic cancer cfDNA is derived from various cell populations including tumour and blood cells.

To see whether tumour DNA could have contributed to altered methylation signal detected by TAPS, we compared DMCs identified by us in cfDNA with differentially methylated CpG sites identified in primary pancreatic tumours by Shen *et al.* (Shen *et al.*, 2018). Briefly, in this study tumour DMCs were identified between tumour and normal tissue and between tumour and peripheral blood mononuclear cells (PBMCs) by RRBS. We found overlaps between cfDNA DMCs and tumour DMCs and performed permutation test to validate their statistical significance. We found statistically significant overlap between hypomethylated DMCs identified in cfDNA and tumour vs. normal tissue comparison (P value = 0, fold enrichment = 12.5) and tumour vs. PBMC (P value = 0, fold enrichment = 11.0). Concordant hypomethylated sites were associated with pro-proliferative genes and oncogenes e.g. SOX18 and KDM2B (Tzatsos *et al.*, 2013; Wang

et al., 2016). Moreover, there was statistically significant overlap between hypermethylated DMCs detected in cfDNA and tumour vs. PBMC (P value = 0, fold enrichment = 2.01). Hypermethylated sites both in cfDNA and primary tumour included tumour suppressor genes and negative growth regulator such as CASP8 and TIMP-2 (Jakubowska *et al.*, 2016; Benzing *et al.*, 2019). Hypomethylated sites exhibited more significant agreement between cfDNA and tumour compared to hypermethylated DMC, however, the exact reason for this observation cannot be determined. It is possible that hypomethylation signal in cfDNA is more tumour-specific while hypermethylation are more likely to be derived from other cell types (e.g. immune cells) rather than malignant cells. Interestingly, we also identified significant overlaps in DMCs with discordant methylation status between tumour and cfDNA including CpG sites hypomethylated in cfDNA but hypermethylated in tumour vs. normal tissue comparison (P value = 0.017, fold enrichment = 1.71) and CpG sites hypermethylated in cfDNA but hypomethylated in tumour vs. normal (P value = 0.029, fold enrichment = 1.39) and tumour vs. PBMC (P value = 0, fold enrichment = 2.09) (Figure 5-12, d-e). This is in line with the study by Chen *et al.* that also detected that fraction of DMCs showed discordant methylation status between cfDNA and tumour (Chen *et al.*, 2020). Inconsistent methylation in cfDNA and tumour can be attributed to domination of methylation signal in plasma DNA by other cell types rather than tumour. This together demonstrated that TAPS could detect methylation signal derived both from tumour and other cell types.

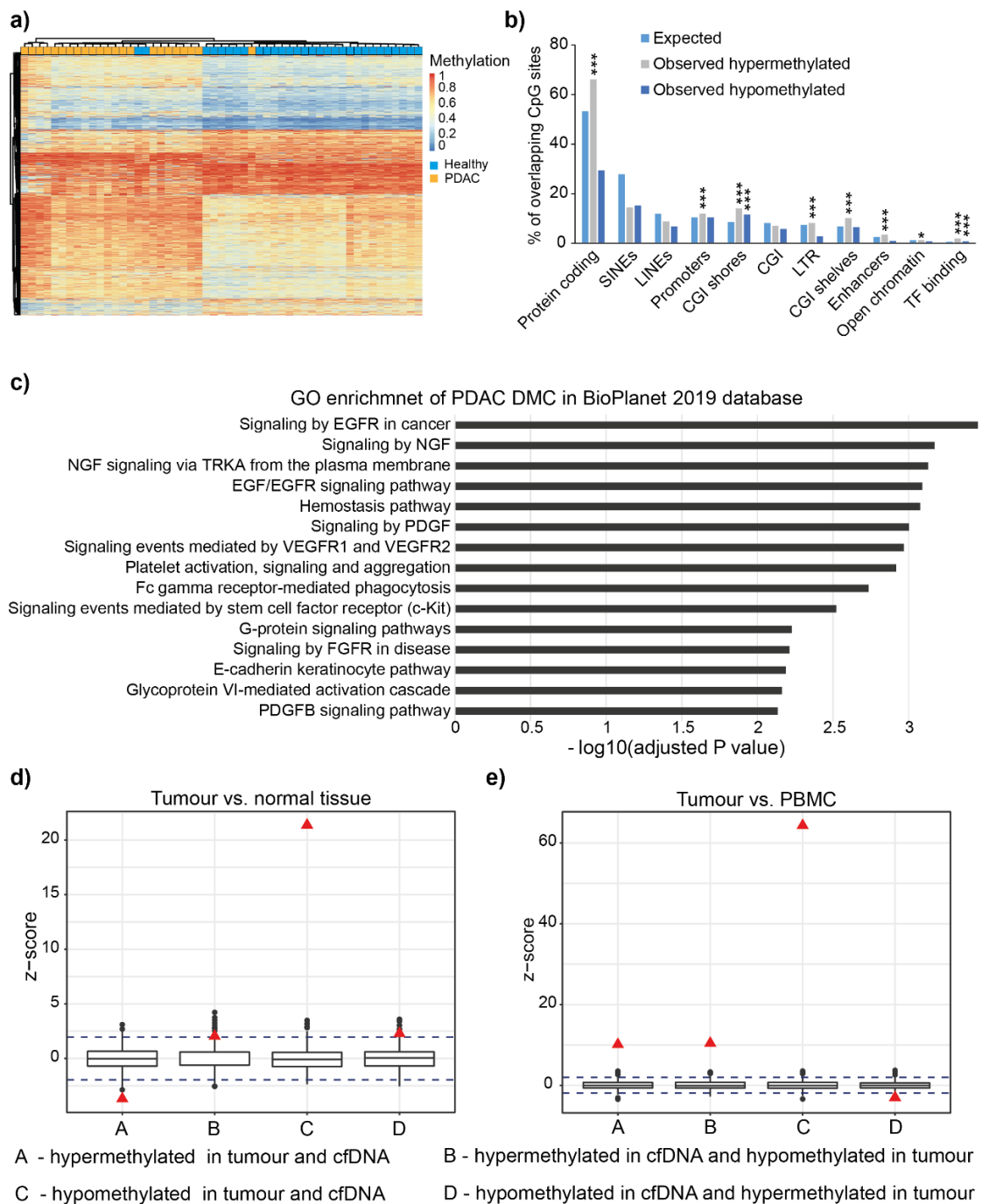


Figure 5-12. Differentially methylated CpG sites in pancreatic cancer cfDNA. (a) Heatmap of 15000 top differentially methylated CpG sites in PDAC cfDNA. (b) Enrichment of cfDNA PDAC DMC in genomic features. Enrichment analysis was performed by Fisher test. Genomic features with significant enrichment of DMC are marked with asterisk (* $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$). (c) Gene Ontology analysis of cfDNA PDAC DMC using Enrichr against BioPlanet 2019 database (Chen *et al.*, 2013; Kuleshov *et al.*, 2016; R. Huang *et al.*, 2019). Top 15 categories selected based on adjusted P value are shown on the graph. (d-e) Overlap between cfDNA PDAC DMC identified by TAPS and DMCs identified based on public RRBS data (Shen *et al.*, 2018) from primary tumour vs. normal tissue comparison (d) and primary tumour vs. PBMCs comparison (e). Boxplots shown expected distribution of overlaps based on 1000 permutations. Z-scores for observed overlaps are represented by triangles.

We performed similar analysis on HCC cfDNA and we found 65516 differentially methylated CpG sites between HCC and controls (P value ≤ 0.01 , methylation difference 0.1), including 35172 hypermethylated sites and 30344 hypomethylated sites (Figure 5-13, a). Similar as observed in PDAC, DMCs were enriched in protein coding regions, CGI shores and shelves and TF binding but not in promoters and LTR. Moreover, hypomethylated DMCs in HCC significantly overlapped with SINE repeats. To note, hypomethylation of SINE was observed in human cancers, however, in the context of HCC loss of methylation in another repetitive element, LINE-1, is better described (Lin *et al.*, 2001; Rauch *et al.*, 2008) (Figure 5-13, b). HCC-specific DMCs were associated with blood cell functions as well as cancer-related pathways with known involvement in HCC pathogenesis e.g. EGFR or netrin-1 signalling pathways (Figure 5-13, c) (Han *et al.*, 2015; S. Z. Huang *et al.*, 2019).

Finally, we compared methylation profiles between HCC and PDAC. Around 14310 DMCs were common between HCC and PDAC DMCs while 40425 CpG sites were differentially methylated between PDAC and HCC cfDNA, including 25935 hypermethylated sites and 14490 hypomethylated sites (Figure 5-14, a). Differentially methylated sites were enriched in regulatory features including promoters and TF binding sites as well as CGI, CGI shores and CGI shelves (Figure 5-14, b). Methylation differences between both cancers could enable to distinguish these cancer types based on cfDNA methylation profile in pan-cancer classification in a future.

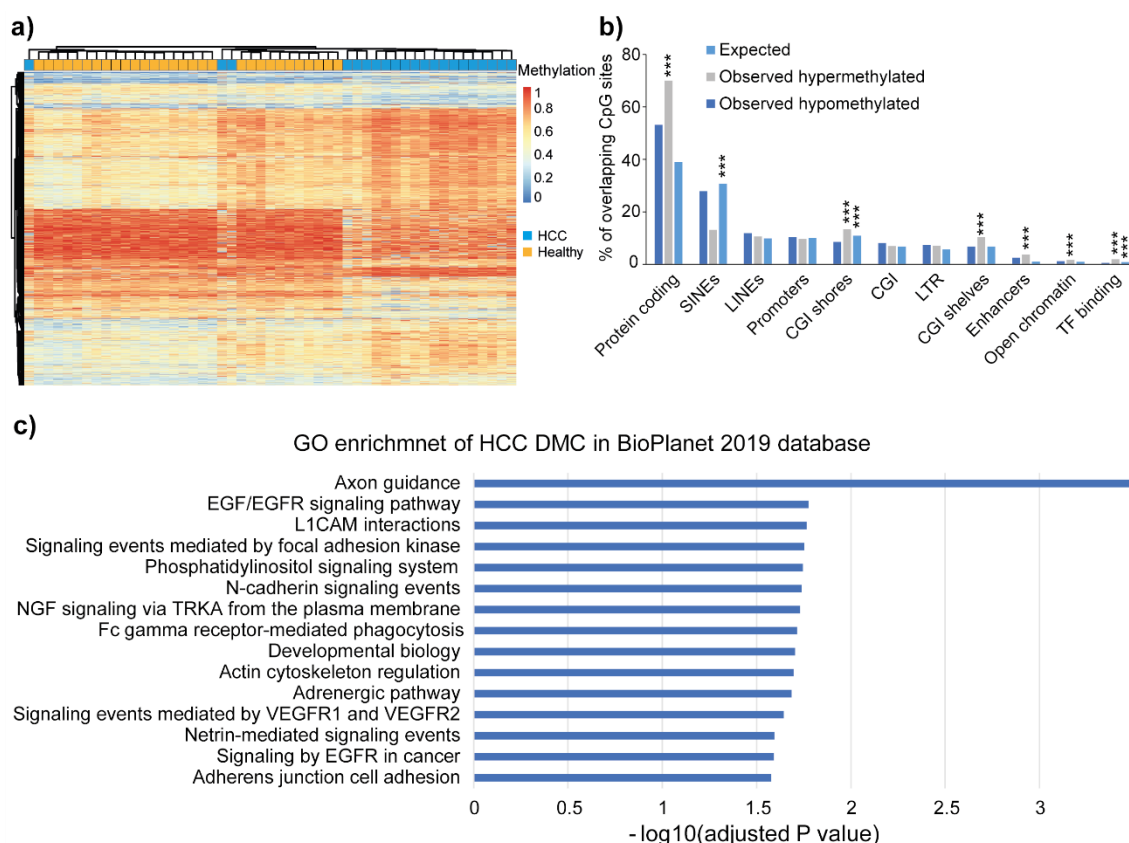


Figure 5-13. Differentially methylated CpG sites in liver cancer cfDNA. (a) Heatmap of 15000 top differentially methylated CpG sites in HCC cfDNA. (b) Enrichment of cfDNA HCC DMC in genomic features. Enrichment analysis was performed by Fisher test. Genomic features with significant enrichment of DMCs are marked with asterisk (* P < 0.05, ** P < 0.01, *** P < 0.001). (c) Gene Ontology analysis of cfDNA HCC DMC using Enrichr against BioPlanet 2019 database (Chen *et al.*, 2013; Kuleshov *et al.*, 2016; R. Huang *et al.*, 2019). Top 15 categories selected based on adjusted P value are shown on the graph.

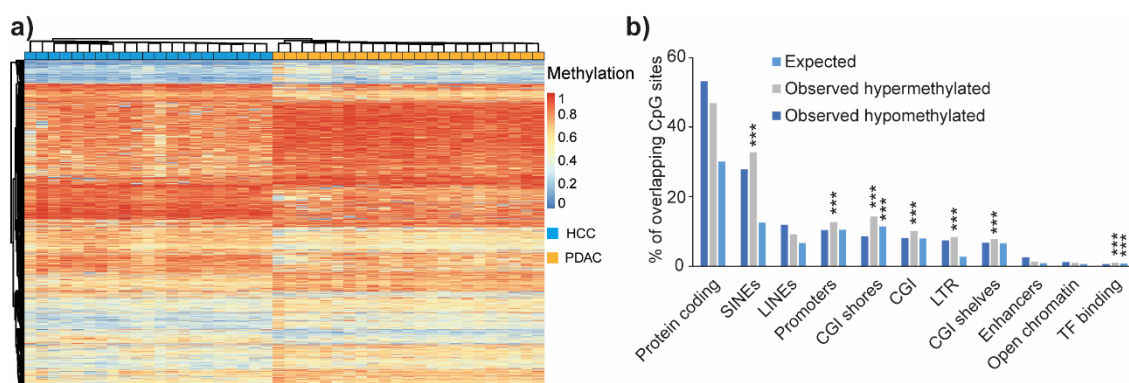


Figure 5-14. Differentially methylated CpG sites between PDAC and HCC cfDNA. (a) Heatmap of 15000 top differentially methylated CpG sites between PDAC and HCC cfDNA. (b) Enrichment of cfDNA PDAC vs. HCC DMC in genomic features. Enrichment analysis was performed by Fisher test. Genomic features with significant enrichment of DMCs are marked with asterisk (* P < 0.05, ** P < 0.01, *** P < 0.001).

Having demonstrated that cfDNA methylomes differ in cancer and non-cancer subjects, we sought to investigate their predictive potential using Leave-One-Out approach for model building. In each training-validation round, one sample was used as an independent model validation set and remaining samples were used as a training set. We identified differentially methylated CpG sites or differentially methylated regions (DMRs) and built a prediction model based on the training set and validated it based on the remaining test sample (Figure 5-15, a). We did not include cirrhosis and pancreatitis samples to build a model, but we used them for an additional validation to evaluate whether our classifiers could reliably discriminate between cancers and pre-malignant conditions.

We tested DMCs and various genomic features for building our predictive models. Results are summarised in Appendix 5 Supplementary Table 5-8.

We achieved the best PDAC prediction based on the top 200 differentially methylated promoters with AUC = 0.96 (65% sensitivity at > 98% specificity) (Figure 5-15, b). Model performance just slightly decreased after including pancreatitis to the validation set (AUC = 0.95, 65% sensitivity at > 98 % specificity), which suggests that promoter methylation has a good ability to discriminate not only between PDAC and healthy controls but also between PDAC and non-malignant conditions. We achieved even better performance for HCC prediction using differentially methylated enhancers (P value < 0.05, methylation difference > 0.05). Our classifier achieved AUC = 0.99 (95% sensitivity at > 98% specificity) without cirrhosis and AUC = 0.98 (95% sensitivity at > 98% specificity) with cirrhosis samples included (Figure 5-15, c). Importantly, our results suggest that developed models were able to correctly classify localised cancers at earlier stages (T1N0M0, T2N0M0) with similar accuracy to more advanced disease suggesting that our approach has a potential for early detection (Figure 5-16, a-d). This together

demonstrates that aberrant methylation in regulatory regions in cfDNA can be a promising feature for cancer prediction with high accuracy, even at the early disease stage. On the other hand, since number of cancer patients with localised disease in our study was low, these findings need to be validated in a bigger cohort.

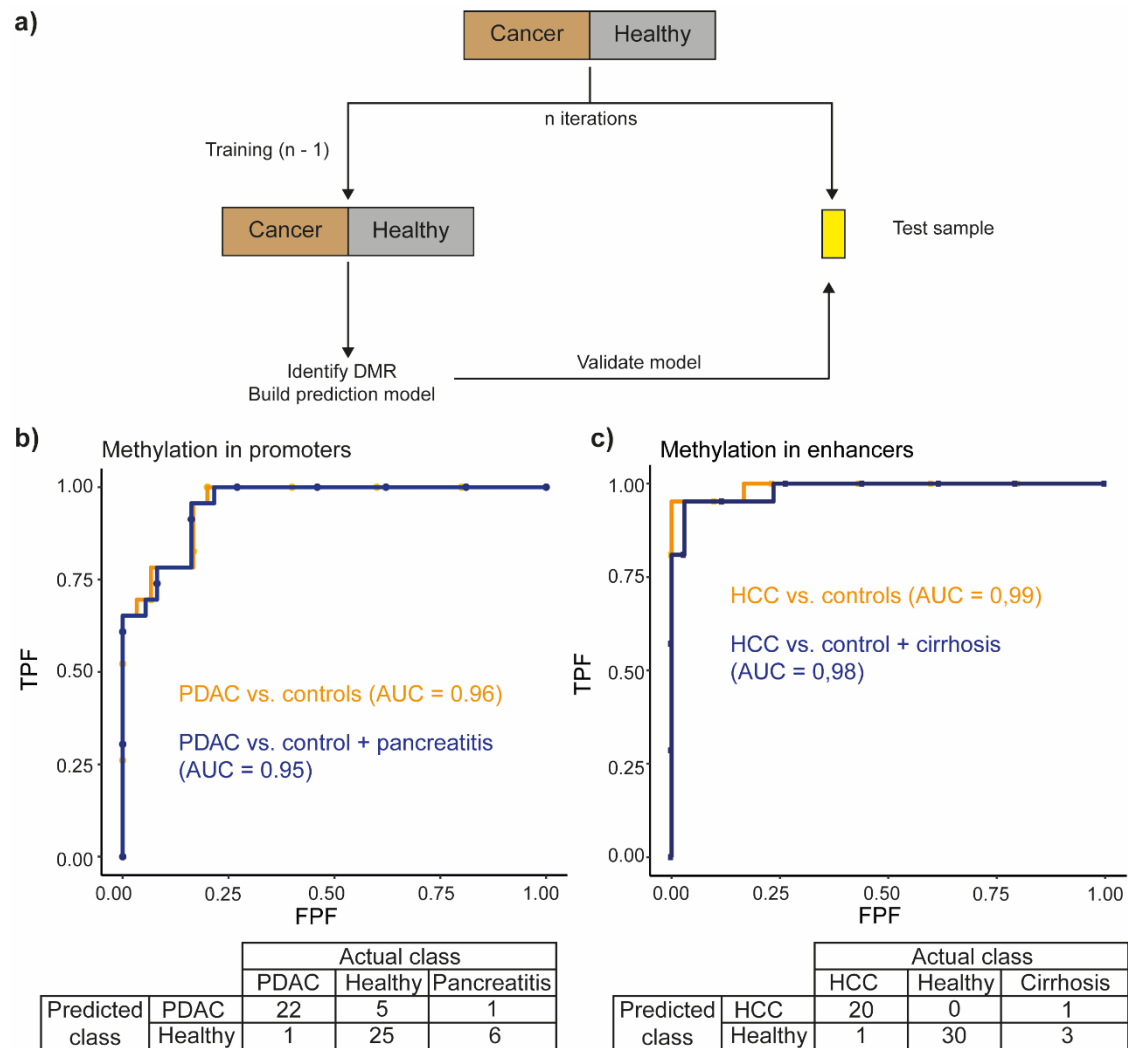


Figure 5-15. Cancer prediction based on cfDNA methylation. (a) Overview of leave-one-out model training and validation protocol. Total number of samples in labelled as n . Training set was defined after excluding a single sample from analysis ($n - 1$). Differentially methylated regions were identified by t-test method (P value < 0.05 , methylation difference > 0.05). All DMR (HCC) or the top 200 DMR (PDAC) were selected for model building. Predictive model was validated on the remaining test sample. Pre-malignant samples were only included in model validation. Final classification of pre-malignant samples was based on the average scores from n models. (b) ROC analysis of methylation classification based on differentially methylated promoters in PDAC and healthy ($n = 53$, PDAC = 23, Healthy = 30, orange plot) and PDAC and healthy + pancreatitis ($n = 60$, PDAC = 23, Pancreatitis = 7, Healthy = 30, navy plot). Number of correct and incorrect predictions in the confusion matrix was determined based on the prediction score threshold corresponding to the best classification accuracy. (c). ROC analysis of methylation classification based on differentially methylated enhancers in HCC and healthy ($n =$

51, HCC = 21, Healthy = 30, orange plot) and HCC, cirrhosis and healthy (n = 55, HCC = 21, Cirrhosis = 4, Healthy = 30). Number of correct and incorrect predictions in the confusion matrix was determined based on the prediction score threshold corresponding to the best classification accuracy.

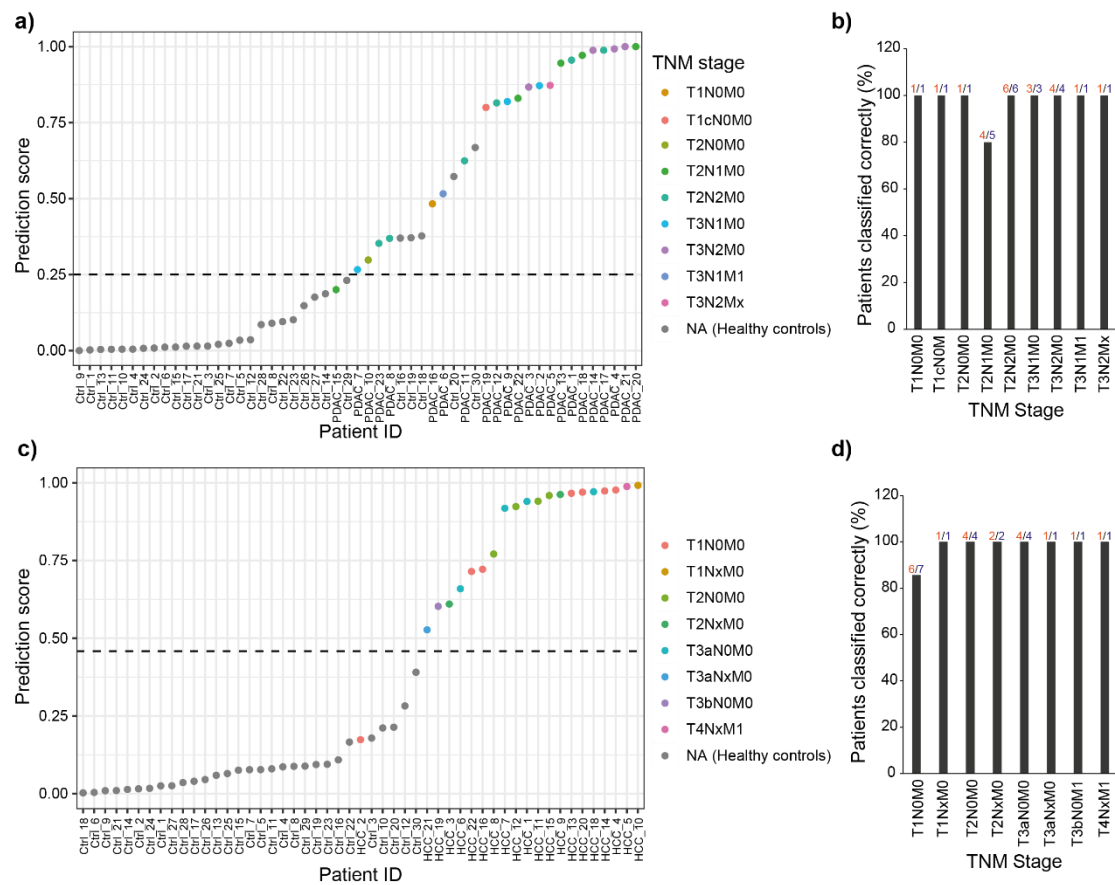


Figure 5-16. PDAC and HCC prediction based on cfDNA methylation in different disease stages. (a) LOO cancer prediction scores for PDAC and non-cancer controls using classifiers trained on the top 200 differentially methylated promoters (P value < 0.05, methylation difference > 0.05) with the lowest P value. Each dot represents an individual patient. Dots representing PDAC patients are coloured by disease stage. The dashed line represents the probability score threshold corresponding to the highest prediction accuracy. Samples with probability score above this threshold were predicted as PDAC. (b) Bar plot showing prediction accuracy of PDAC patients depending on TNM stage in the model trained based on the top 200 differentially methylated promoters. Numbers above bars show the number of correctly predicted patients (orange) and the total number of patients at the particular disease stage (navy). (c) LOO cancer prediction scores for HCC and non-cancer controls using classifiers trained on the differentially methylated enhancers (P value < 0.05, methylation difference > 0.05). Each dot represents an individual patient. Dots representing HCC patients are coloured by disease stage. The dashed line represents the probability score threshold corresponding to the highest prediction accuracy. Samples with probability score above this threshold were predicted as HCC. (d) Bar plot showing prediction accuracy of HCC patients in the model trained based on differentially methylated enhancers depending on TNM stage. Numbers above bars show the number of correctly predicted patients (orange) and the total number of patients at the particular disease stage (navy).

5.4 CfDNA fragmentation profiles are preserved in TAPS and have predictive potential

One of the most important advantages of TAPS is its non-damaging character and direct conversion of modified cytosines that leaves most of nucleotides in the genome intact. Therefore, TAPS data could provide additional genetic information besides DNA methylation. One of potentially important non-epigenetic cfDNA feature is fragmentation profile. Recent studies have demonstrated that cfDNA fragmentation patterns differ in cancer patients compared to healthy (Jiang *et al.*, 2015; Mouliere *et al.*, 2018; Cristiano *et al.*, 2019). However, fragmentation information can be mainly retrieved from standard, paired-end whole genome sequencing and fragmentation profiles may be affected in methylation sequencing methods involving bisulfite treatment (Erger *et al.*, 2020). Since TAPS does not affect DNA length, we sought to investigate whether our sequencing can provide reliable information about cfDNA fragmentation profiles and whether this information can be used for disease prediction.

We first confirmed that cfDNA fragmentation profiles retrieved from TAPS sequencing are consistent with expected cfDNA fragmentation pattern with the major peak at ~ 167 bp and secondary peak at ~ 320 bp. Smaller peaks below the major peak correspond to the nucleosome fragmentation pattern (Figure 5-17, a). In agreement with literature we found that cancer patients have statistically significant higher proportion of short cfDNA fragments below 150 bp (HCC: $P = 6.871\text{e-}06$, PDAC: $P = 0.006731$) (Mouliere *et al.*, 2018) (Figure 5-17, b). Moreover, we found significantly lower fraction of long cfDNA fragments (310-500 bp) in PDAC ($P = 1.263\text{e-}06$) and HCC ($P = 2.627\text{e-}07$) compared to non-cancer controls. Frequency of long cfDNA fragments (310-500 bp) in 10 bp range length bins could separate PDAC and HCC from controls in unsupervised analysis (Figure 5-17, c-d)

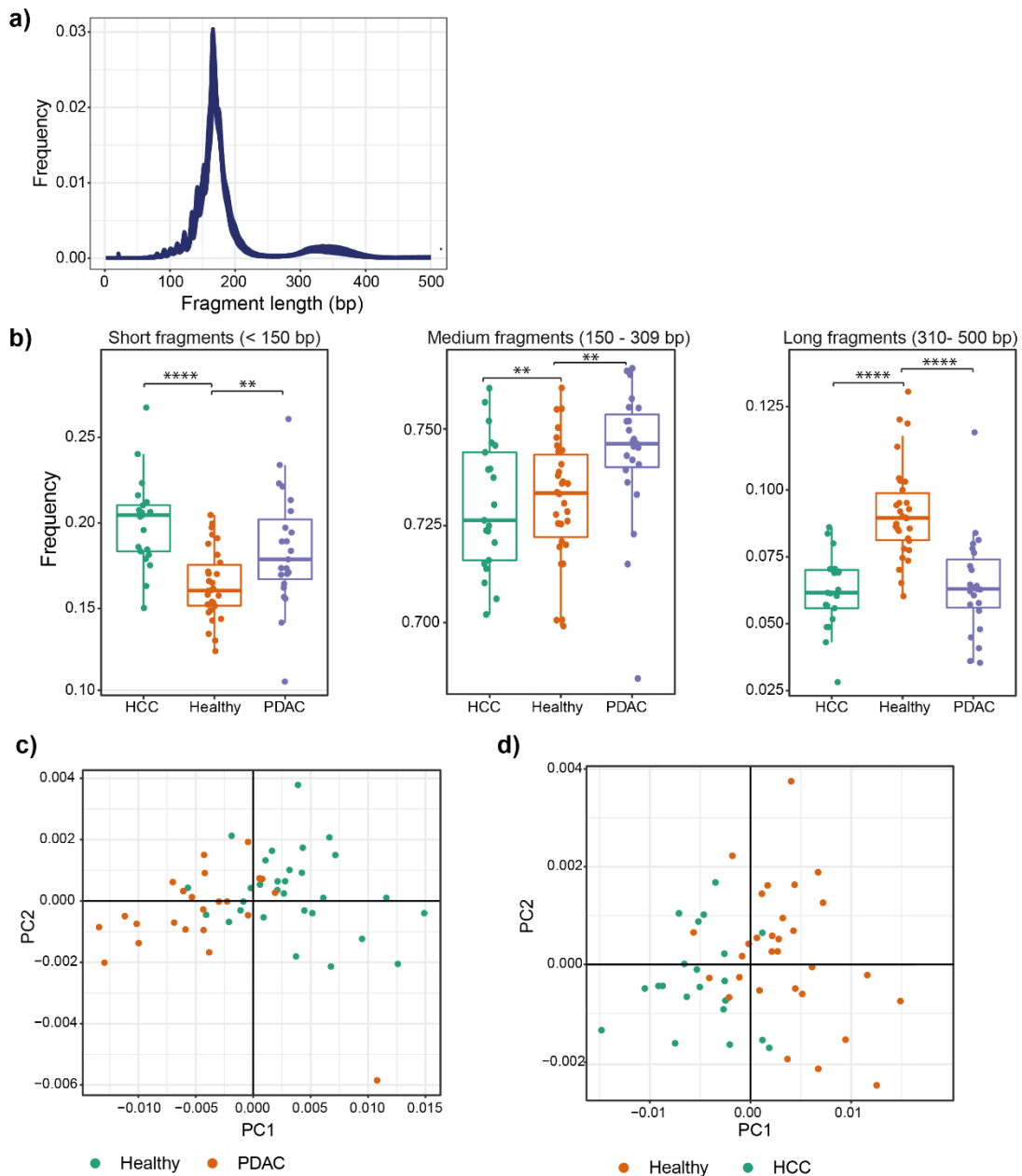


Figure 5-17. CfDNA fragmentation profiles in cancer and healthy. (a) Fragment size distribution of 30 healthy cfDNA samples. Frequency was calculated as number of fragments of particular length divided by total number of fragments. (b) Proportion of short cfDNA fragments (below 150 bp), medium cfDNA fragments (150-309 bp) and long cfDNA fragments (310-500 bp) in non-cancer controls, PDAC and HCC. The Kruskal-Wallis test was performed to test differences in fragment size distribution between groups. Statistically significant differences are marked with asterisk (* $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, **** $P < 0.0001$). PCA analysis of long fragments frequency for (c) PDAC and healthy and (d) HCC and healthy.

We tested predictive potential of cfDNA fragmentation in a long cfDNA range (310-500 bp) to distinguish between cancer and controls. Fragmentation profiles performed worse at PDAC prediction compared to promoter methylation. Our fragmentation-based

classifier achieved AUC = 0.89 for PDAC vs. healthy comparison. Model performance after including pancreatitis to the validation set was very similar (AUC = 0.88) (Figure 5-18, a).

Similar to PDAC, fragmentation exhibited good potential to discriminate between HCC and controls, however, overall prediction was worse compared to methylation-based models. We achieved AUC = 0.91 for HCC vs. healthy comparison. Prediction accuracy decreased after including cirrhosis into the test group (AUC = 0.84). Notably, all cirrhosis samples were misclassified as HCC suggesting that fragmentation profiles are worse than methylation at discriminating between liver cancer and liver chronic conditions (Figure 5-18, b).

Predictors trained based on cfDNA fragmentation profiles were able to correctly classify most PDAC and HCC patients, however, they missed more localised cancers compared to methylation-based models. It suggests that cfDNA fragmentation profiles may have lower sensitivity to detect early stage disease compared to methylation, however, due to the low number of patients in our study, these findings need to be confirmed in a bigger cohort (Figure 5-19, a-d).

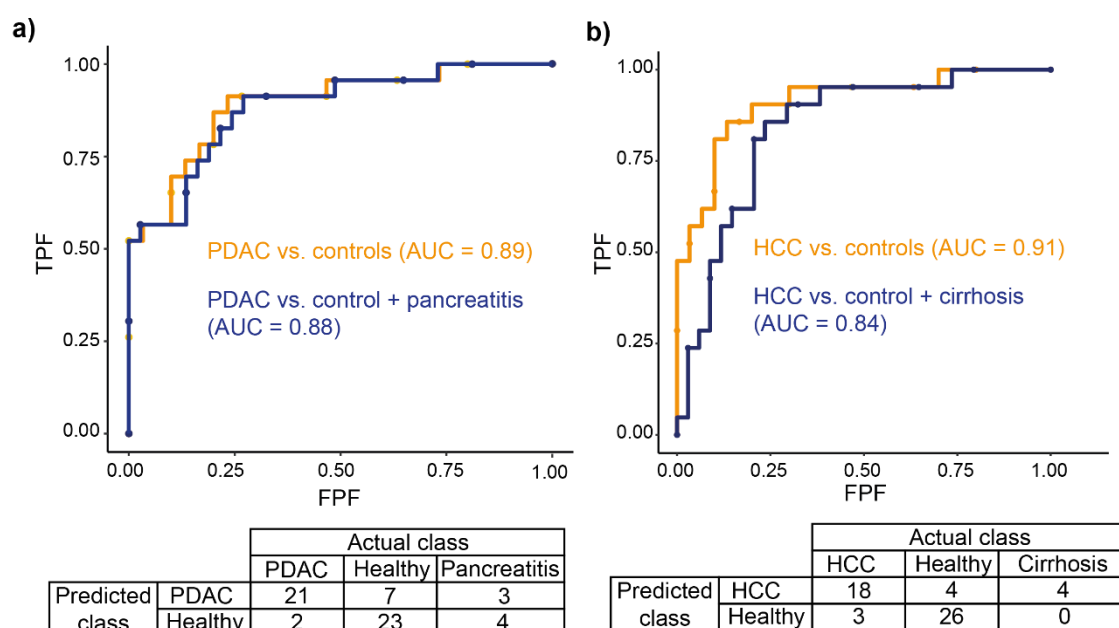


Figure 5-18. (a) ROC analysis of classification based on cfDNA fragmentation in long cfDNA range (310-500 bp) for PDAC vs. healthy (n = 53, PDAC = 23, Healthy = 30, orange plot) and PDAC vs. healthy + pancreatitis (n = 60, PDAC = 23, Pancreatitis = 7, Healthy = 30, navy plot). Number of correct and incorrect predictions in the confusion matrix was determined based on the prediction score threshold corresponding to the best classification accuracy. (b). ROC analysis of classification based on cfDNA fragmentation in long cfDNA range (310 – 500 bp) for HCC and healthy (n = 51, HCC = 21, Healthy = 30, orange plot) and HCC vs. cirrhosis and healthy (n = 55, HCC = 21, Cirrhosis = 4, Healthy = 30). Number of correct and incorrect predictions in the confusion matrix was determined based on the prediction score threshold corresponding to the best classification accuracy.

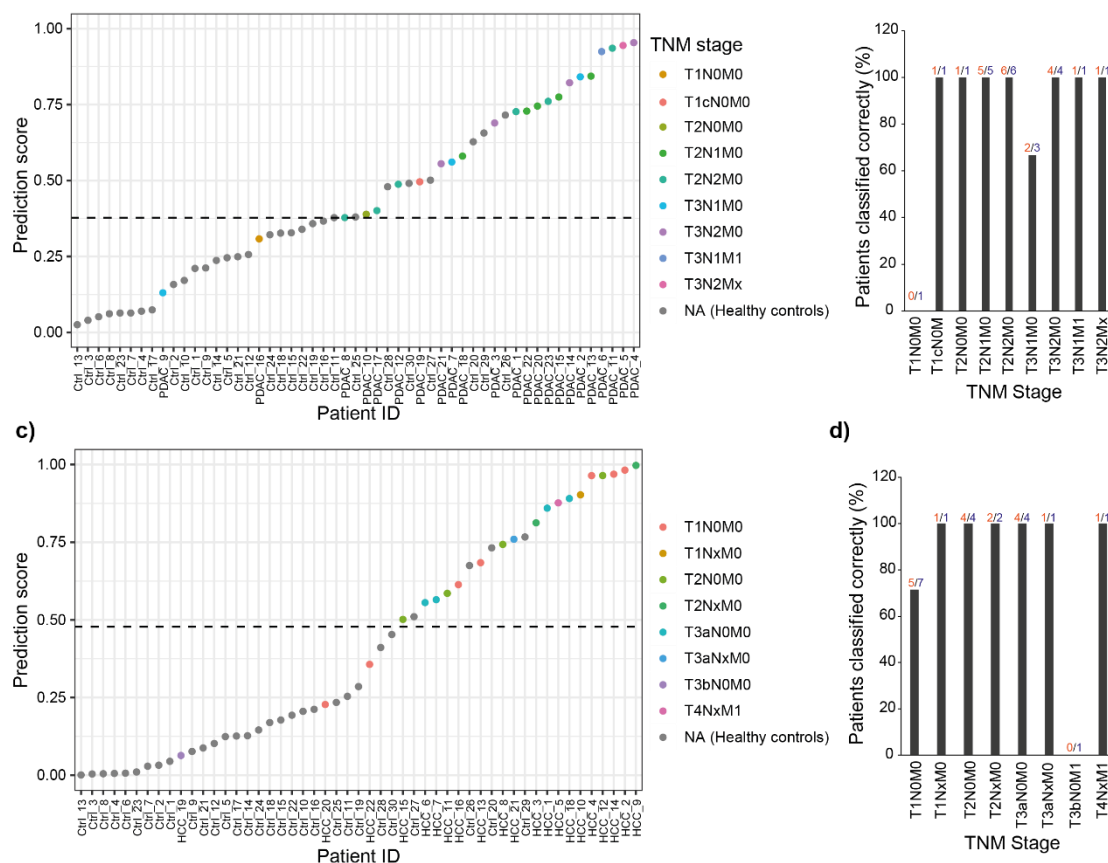


Figure 5-19. PDAC and HCC prediction based on based on cfDNA fragmentation in long cfDNA range (310-500 bp) in different disease stages. (a) LOO cancer prediction scores for PDAC and non-cancer controls using classifiers trained based on cfDNA fragmentation in long cfDNA range (310-500 bp). Each dot represents an individual patient. Dots representing PDAC patients are coloured by disease stage. The dashed line represents the probability score threshold corresponding to the highest prediction accuracy. Samples with probability score above this threshold were predicted as PDAC. (b) Bar plot showing prediction accuracy of PDAC patients depending on TNM stage in the model trained based on cfDNA fragmentation in long cfDNA range (310-500 bp). Numbers above bars show the number of correctly predicted patients (orange) and the total number of patients at the particular disease stage (navy). (c) LOO cancer prediction scores for HCC and non-cancer controls using classifiers trained based on cfDNA fragmentation in long cfDNA range (310-500 bp). Each dot represents an individual patient. Dots representing HCC patients are coloured by disease stage. The dashed line represents the probability score threshold corresponding to the highest prediction accuracy. Samples with probability score above this threshold were predicted as HCC. (d) Bar plot showing prediction accuracy of HCC patients in the model trained based on cfDNA fragmentation in long cfDNA range (310-500 bp) depending on TNM stage. Numbers above bars show the number of correctly predicted patients (orange) and the total number of patients at the particular disease stage (navy).

To summarise, we demonstrated that TAPS method preserves information about cfDNA fragmentation and that fragmentation profiles can predict cancer with good accuracy. On the other hand, fragmentation-based prediction had lower specificity compared with methylation-based classifiers, especially for non-malignant, chronic conditions.

Nevertheless, we proved that besides methylation TAPS data can provide other cfDNA characteristics enabling simultaneous, whole-genome methylation and fragmentation analysis from a single sequencing assay.

5.5 Discussion

Cancer detection based on information found in cfDNA is considered as a promising approach for development of novel diagnostic tools. Circulating DNA fragments contain multiple layers of information including mutations (Cohen *et al.*, 2018), fragmentation profiles (Mouliere *et al.*, 2018; Cristiano *et al.*, 2019) and methylation changes (Shen *et al.*, 2018; M. C. Liu *et al.*, 2020). Differences between cfDNA in cancer and healthy can be used for non-invasive disease detection.

Recent research has demonstrated that cfDNA methylation enables sensitive and specific detection of cancer (Shen *et al.*, 2018; M. C. Liu *et al.*, 2020; Nassiri *et al.*, 2020). Moreover, genetic changes including cfDNA fragmentation profiles can also distinguish cancer from healthy (Mouliere *et al.*, 2018; Cristiano *et al.*, 2019). Analysis of multiple cfDNA features in a single assay could enable comprehensive and cost-effective approach for cfDNA sequencing. Unfortunately, most of genetic information such as mutations or fragmentation profiles cannot be retrieved from sequencing data produced by current gold standard for whole-genome methylation sequencing, BS-seq. Harsh chemical treatment involved in this method may affect natural cfDNA fragmentation patterns and is not suitable for simultaneous analysis and methylation and fragmentation signatures (Erger *et al.*, 2020). Moreover, bisulfite treatment converts unmodified cytosines that constitute of around 20% of the genome. Since base-change is introduced in a significant fraction of the genome, detection of mutations in bisulfite data is challenging.

We developed a new, bisulfite-free method for single-base resolution methylation sequencing called TAPS. We adopted this method for whole-genome methylation sequencing of cfDNA and demonstrated that TAPS can provide sensitive and reliable detection of per base methylation and in many aspects outperforms other epigenetic sequencing methods.

To the best of our knowledge, this is along to EM-seq, the only bisulfite-free, base-resolution methylation sequencing method already validated in cfDNA. To note, EM-seq cfDNA libraries exhibited disappointingly low accuracy in detecting modified cytosines, which resulted in over 20% false negative rate (Erger *et al.*, 2020). In regard to conversion, TAPS performed significantly better and achieved 96% of true positives and only 0.28% of false positive rate.

Another great advantage of TAPS is its non-damaging, mild character. It allowed us to prepare sequencing libraries from as low as 10 ng of cfDNA extracted from 2-4 mL of plasma. This is particularly important since DNA level in a circulation is low (Wan *et al.*, 2017). Although cfDNA bisulfite sequencing is technically feasible, it requires significantly more cfDNA input to produce good quality and complexity libraries. For instance, CCGA study testing predictive potential of cfDNA methylation based on bisulfite sequencing required up to 75 ng of cfDNA for deep targeted sequencing (M. C. Liu *et al.*, 2020). This level of DNA material requires significant plasma volume and might not be possible to obtain from some patients.

TAPS outperforms whole-genome BS-seq also in mapping efficiency. In our study, we generated over 360M reads per library. We achieved high unique mapping rate over 94% and in average 76% of reads were uniquely aligned and deduplicated. To compare, Allen Chan *et al.* prepared whole genome BS-seq cfDNA libraries and sequenced them to 163 million reads. BS-seq libraries exhibited worse mapping rate with 78.4% of uniquely

mapped reads among which 73.3% were non-duplicated. It resulted in only 57% of unique, deduplicated reads and over 40% of sequenced reads were redundant (Allen Chan *et al.*, 2013).

This demonstrates that cfDNA TAPS yields more informative reads and thus is more cost-effective compared to BS-seq. Improved mapping rate observed for TAPS is likely due to direct conversion of modified cytosines and high sequence complexity close to whole genome sequencing. In turn, its non-damaging character retains more amplifiable template and therefore TAPS libraries can be amplified by less PCR cycles resulting in lower duplication rate.

Another cfDNA methylation sequencing method, cfMeDIP-seq, is able to generate sequencing libraries from 10 ng of cfDNA, similar to TAPS, however it did not provide fully quantitative, base-resolution information about methylation (Shen *et al.*, 2018; Nassiri *et al.*, 2020).

Finally, we confirmed that other cfDNA features can be retrieved from TAPS data besides methylation. We demonstrated that in contrast to BS-seq, TAPS retains natural fragmentation patterns such as position and length of the major cfDNA peaks and 10 bp periodic oscillations similar to whole genome sequencing, which enables simultaneous methylation and fragmentation analysis (Mouliere *et al.*, 2018; Erger *et al.*, 2020). Since only modified cytosines are converted in TAPS protocol and ~95% of nucleotides remain intact, detection of mutations from TAPS data should be feasible in a future after novel computational tools have been developed. This would allow for fully comprehensive cfDNA profiling in a single sequencing assay.

Taken together, TAPS offers numerous advantages to current techniques available for genome wide methylation sequencing and has potential to facilitate and improve integrated epigenetic and genetic analysis of cfDNA.

The ultimate goal of cfDNA TAPS development was its application to cfDNA to confirm that this method can provide information with diagnostic potential. We applied TAPS to the cfDNA samples from HCC, PDAC and non-cancer controls and demonstrated that even at non-metastatic stage, cancer patients exhibited distinct methylation and fragmentation profiles with predictive potential. Our analysis showed that methylation in regulatory elements allowed for very accurate cancer prediction in our dataset. Namely, we were able to accurately classify HCC with $AUC = 0.99$ based on enhancers methylation and PDAC with $AUC = 0.96$ based on methylation in promoters. Fragmentation-based predictors performed worse with $AUC = 0.91$ for HCC and $AUC = 0.89$ for PDAC. Importantly, fragmentation-based predictors were characterised by lower specificity compared to the classifiers developed on methylation data and predicted more healthy or pre-malignant patients as cancer.

We demonstrated that methylation in regulatory regions provided by TAPS enabled accurate prediction of HCC with high $AUC = 0.99$. This performance is already better than other published methylation biomarkers ($AUC = 0.81-0.96$) (Allen Chan *et al.*, 2013; Cai *et al.*, 2019; Sun *et al.*, 2019; Kotoh *et al.*, 2020). Our models also showed improved performance compared with clinically used serum protein biomarker AFP that has estimated $AUC = 0.80$ or 63% sensitivity at 90.6% specificity (Trevisani *et al.*, 2001; Kisiel *et al.*, 2019). To note, our prediction was similar to multianalyte panel, CancerSEEK, that combines mutations with protein biomarkers and achieved 100% sensitivity for HCC detection (Cohen *et al.*, 2018).

Despite the promising results, our findings need to be validated in the future on an independent set of samples to confirm their translational potential. Important aspect that needs to be considered is discrimination between HCC and chronic liver diseases such as liver cirrhosis. Cirrhosis is one of the biggest risk factors for HCC and it is indication for

extensive medical surveillance. HCC detection based on currently available diagnostic methods including imaging techniques and AFP is affected by presence of cirrhotic lesions often leading to misdiagnosis (Trevisani *et al.*, 2001; Kansagara *et al.*, 2014; Pascual *et al.*, 2019). Therefore, it is crucial that novel diagnostic tools are able to distinguish between malignant and benign liver condition with high confidence. At this stage, our classifier showed great potential to detect HCC among non-cancer controls and very promising ability to distinguish cirrhosis from HCC with all cirrhotic patients classified as healthy. It is important to note that our current cohort included only 4 cirrhotic patients and our results have to be validated and optimised for HCC and cirrhosis discrimination in a bigger cohort. Ideally, a future cohort should include more cirrhosis patients as well as with multiple HCC subcohorts designed based on disease aetiology. Such a study would facilitate a complex comparison between chronic liver diseases and cancer groups and ultimately allow for a perfect discrimination between benign and malignant liver-derived signals in cfDNA for improved specificity.

Another cancer type we included into our study was PDAC. Our classifier was able to predict PDAC among non-cancer controls with AUC = 0.96 based on promoter methylation. This performance is comparable or even better than clinically used biomarker CA19-9. Depending on study, CA19-9 sensitivity and specificity are estimated as 47-79% and 82-88%, respectively (Goonetilleke and Siriwardena, 2007; Cao *et al.*, 2016; Ngamruengphong and Canto, 2016). For comparison, our assay sensitivity ranges from 78-96% at the same sensitivities. Another study estimated that CA19-9 can detect PDAC among healthy cohort with AUC = 0.76-0.78 and among pancreatitis patients with AUC = 0.73-0.77 (Haab *et al.*, 2015). Our classifier performed better to separate PDAC from non-cancer controls (AUC = 0.96) and PDAC from pancreatitis (AUC = 0.95). This

suggests that methylation-based approaches can more accurately discriminate between benign and malignant pancreas diseases than traditional, protein-based biomarkers.

TAPS data provided information that enabled us to detect PDAC with similar or even better performance not only to CA19-9 but also compared to some recently published biomarkers. For instance, CCGA study that achieved around 60-80% sensitivity for stage I and II PDAC. On the other hand, CCGA study used higher sample size than our pilot study and its performance was measured based on target sites pre-selected based on previous study (M. C. Liu *et al.*, 2020). Other methylation-based approaches detected PDAC with AUC between 0.86-0.92 (Henriksen *et al.*, 2016; Shen *et al.*, 2018).

As demonstrated our methylation-based classifiers could accurately predict cancer and distinguish them from control group even though the control group did not include completely healthy individuals. Control samples were collected from Endoscopy clinic and they included patients exhibiting gastro-intestinal symptoms, some of them suffering from chronic, inflammatory conditions such as Crohn disease or colitis. Chronic inflammatory conditions in the controls could alter non-cancer methylation and fragmentation patterns, contributing to an increased noise in the control group and affect model performance. However, even with this challenging controls, our models still accurately detected cancer. Our classifiers could potentially perform even better with the bigger sample size where effects of comorbidities are usually more diluted or if control group included individuals with comorbidities more characteristic for a particular age group and not enriched for one, particular type of chronic, systemic diseases.

Great promise of our assay was its ability to predict cancer with high accuracy even in the cohort that mostly included non-metastatic cancer patients. It is especially important for PDAC, as pancreas contribution to cfDNA is relatively low and it is possible that tumour fraction in these patients was low (Mouliere *et al.*, 2018). However, our assay

was still able to correctly predict most of PDAC samples. Possibly whole-genome, tumour-naïve approach for methylation analysis in cfDNA contributed to the sensitive cancer prediction. Since our analysis suggests that cfDNA methylomes in PDAC and HCC are different and methylation signatures are tissue-specific, a multi-class predictor for simultaneous HCC and PDAC detection can be developed in a future. To conclude, cfDNA TAPS provides a tool for multidimensional analysis of cfDNA in healthy and cancer and provides information with predictive potential. Developed methylation-based classifiers were able to detect HCC and PDAC with high specificity and sensitivity and outperformed cancer prediction based on fragmentation patterns. Although this chapter showed only methylation and fragmentation analysis, other cfDNA characteristic including mutations, CNV or tissue-of-origin can be potentially extracted from TAPS data.

5.6 Concluding remarks

Here we introduced TAPS method for whole-genome, base-resolution methylation sequencing of cfDNA. We demonstrated that TAPS offers significant, technological improvements compared to other methods available for cfDNA methylation sequencing and provides per base methylation information from as low as 10 ng of cfDNA. We also showed that apart from methylation, TAPS retains reliable information about cfDNA fragmentation profiles. Methylation signatures in regulatory features enabled for sensitive and specific detection of HCC and PDAC.

Future work will focus on validation of findings of this study on a bigger, independent cohort and on development of multicancer predictors.

5.7 Materials and Methods

5.7.1 Preparation of carrier DNA and spike-in controls

Carrier DNA was prepared by PCR amplification of the pNIC28-Bsa4 plasmid (Addgene, cat. no. 26103) in a reaction containing 1 ng DNA template, 0.5 μ M primers and 1 $\times$ Phusion High-Fidelity PCR Master Mix with HF Buffer (Thermo Scientific). Primer sequences are listed in Table 5-2.

Table 5-2. Primer sequences used to amplify carrier DNA.

Template	Primer	Sequence (5' to 3')
Carrier 2 kb	2 kb-F	AGGCAACTTTATGCCCATGCAA
	2 kb-R	CCAAGGGGTTATGCTAGTTATTGC

Spike-in control including CpG methylated lambda DNA, 2 kb unmodified control and 160 mer bp oligo were prepared as described in section 4.11.2. CpG methylated lambda DNA, carrier DNA and 2 kb unmodified control were fragmented by Covaris M220 (Peak Incident Power – 50 W, Duty Factor - 20%, Cycles per Burst (cpb) – 200, time – 150 s) and size-selected on 0.9-1.2 $\times$ AMPure XP beads to select for 150-250 bp fragments.

5.7.2 Preparation of mTET1CD

mTET1CD was expressed and purified as described in section 4.12.5. Enzyme activity was measured as described in section 4.11.8.

5.7.3 Preparation of sequencing adapters

Short Illumina adapters were prepared as described in section 4.11.11.

5.7.4 mTet1CD oxidation

DNA was incubated in a 50 μ l reaction containing 50 mM HEPES buffer (pH 8.0), 100 μ M ammonium iron(II) sulfate, 1 mM α -ketoglutarate, 2 mM ascorbic acid, 1 mM dithiothreitol, 100 mM NaCl, 1.2 mM ATP and 4 μ M mTet1CD for 80 min at 37°C. After that, 0.8 U of Proteinase K (New England Biolabs) were added to the reaction mixture

and incubated for 1 h at 50°C. The product was cleaned up on Bio-Spin P-30 Gel Column (Bio-Rad) and 1.8× AMPure XP beads following the manufacturer's instruction.

5.7.5 Pyridine borane reduction

Oxidized DNA in 35 µl of water was reduced in a 50 µl reaction containing 600 mM sodium acetate solution (pH 4.3) and 1 M pyridine borane for 16 h at 37°C and 850 r.p.m. in an Eppendorf ThermoMixer. The product was purified using Zymo-Spin columns.

5.7.6 Preparation of low input TAPS libraries

mESC genomic DNA prepared as described in section 4.11.12 for whole-genome TAPS was used for low-input whole-genome TAPS. Briefly, samples containing 100 ng, 10 ng and 1 ng of mESC genomic DNA were used for an end-repair and A-tailing reaction and ligated to Illumina Multiplexing adapters with KAPA HyperPlus kit according to the manufacturer's protocol. Ligated samples were then oxidized with mTet1CD once and then reduced with pyridine borane according to the protocols described in sections 5.7.4 and 5.7.5. Converted libraries were amplified with KAPA Hifi Uracil Plus Polymerase for 5 cycles (100 ng), 8 cycles (10 ng) and 13 cycles (1 ng) and cleaned up on 1× AMPure XP beads.

CfDNA TAPS samples were prepared from 10 ng and 1 ng of cell-free DNA sample as described above for whole-genome TAPS. Briefly, cfDNA samples were used for an end-repair and A-tailing reaction and ligated to Illumina Multiplexing adapters with KAPA HyperPlus kit according to the manufacturer's protocol. Ligated samples were then oxidized with mTet1CD once and then reduced with pyridine borane according to the protocols described in sections 5.7.4 and 5.7.5. Converted libraries were amplified with KAPA Hifi Uracil Plus Polymerase for 7 cycles (10 ng), 13 cycles (1 ng) and cleaned up on 1× AMPure XP beads.

5.7.7 Collection and preparation of cfDNA samples

Non-cancer control patients were recruited from John Radcliffe hospital (Ethical approvals IDs 16/YH/0247 and 18/WM/0237). Pancreatitis patients were recruited John Radcliffe hospital. The study was approved by Oxfordshire REC-A (10/H0604/51) and is registered on the UK NIHR portfolio as study number 10776. All patients were recruited from Churchill hospital (Oxford) consented for this study via the Oxford Radcliffe biobank (09/H0606/5+5, project: 19/A177). HCC and cirrhosis samples collections were REC approved (Ethical approval 2/NE/0395, IRAS project ID = 116370).

Whole blood (5-20 mL) from non-cancer controls, pancreatitis and PDAC was collected into EDTA tubes and kept on ice. Plasma was separated within 1-4 hours after blood collection. Briefly, blood was transferred to Falcon™ 15 mL Conical Centrifuge Tubes and centrifuged at 1600 xg for 10 min at 4°C. Subsequently, top liquid layer was collected carefully without disturbing cellular pellet, transferred to DNA LoBind® Tubes (Eppendorf) and centrifuged at 16000 xg at 10 min at 4°C. Collected plasma was then divided into 1 mL aliquots and stored in DNA LoBind® Tubes (Eppendorf) at -80°C until cfDNA isolation.

cfDNA was extracted from 1-4 mL of plasma using Qiaamp Circulating Nuclei Acid Kit (Qiagen) and eluted in EB buffer. cfDNA was quantified by Qubit Fluorometer (Life Technologies). Integrity and size distribution of cfDNA were assessed based on electrophoretic separation and visualisation on 2% E-Gel EX Agarose Gel (Invitrogen). All cfDNA samples included in the study exhibited distinct sharp band at ~ 170 bp and did not show any signs of DNA degradation or contamination with genomic DNA. Isolated cfDNA was stored in DNA LoBind® Tubes (Eppendorf) at -20°C.

HCC and cirrhosis plasma samples (1.5-2 mL) were shipped frozen from Newcastle University. Blood and plasma were collected as described above for non-cancer controls, pancreatitis and PDAC patients. cfDNA was extracted from plasma using Qiaamp Circulating Nuclei Acid Kit (Qiagen) and eluted in EB buffer. cfDNA was quantified by Qubit Fluorometer (Life Technologies). Integrity and size distribution of cfDNA were assessed based on electrophoretic separation and visualisation on 2% E-Gel EX Agarose Gel (Invitrogen). Isolated cfDNA was stored in DNA LoBind® Tubes (Eppendorf) at -20°C.

5.7.8 cfDNA TAPS with carrier DNA

10 ng of cfDNA were spiked-in with 0.1% CpG-methylated lambda DNA, 0.01% unmodified 2 kb control and 0.05% synthetic oligo control with N5mCNN and N5hmCNN and used for an end-repair and A-tailing reaction and ligated to Illumina Multiplexing adapters with KAPA HyperPlus kit according to the manufacturer's protocol. Subsequently 100 ng of carrier DNA were added to ligated libraries and samples were double-oxidized with mTet1CD and reduced with pyridine borane according to the protocols described in sections 5.7.4 and 5.7.5. Converted libraries were amplified with KAPA Hifi Uracil Plus Polymerase for 6 or 7 cycles (10 ng) and cleaned up on 1× AMPure XP beads. Three pilot cfDNA libraries prepared by carrier DNA TAPS protocol were paired-end 80 bp sequenced on a NextSeq 500 sequencer (Illumina) using one NextSeq High Output kit with 0.5% PhiX control library spike-in. Final cfDNA cohort (96 libraries) were paired-end 150 bp sequenced on a NovaSeq 6000 sequencer (Illumina).

5.7.9 TAPS mapping and pre-processing

Pilot cfDNA sequencing

Paired-end reads were downloaded from Illumina BaseSpace and subsequently quality trimmed with Trim Galore! v0.6.0. Read pairs where at least one read was shorter than 35 bp after trimming were removed. Trimmed reads were mapped to a genome combining spike-in sequences, and the human reference genome (GRCh38) using asTair v3.2.4 (<https://bitbucket.org/bsblabludwig/astair/>). PCR duplicates were called using Picard v2.20.2 MarkDuplicates (<http://broadinstitute.github.io/picard/>) Modified bases were detected using deduplicated and uniquely mapped reads using asTair v3.2.4 (<https://bitbucket.org/bsblabludwig/astair/>).

cfDNA final cohort

Raw sequenced reads were processed with trim_galore (version 0.6.2) to trim adapter and low-quality bases with the following parameters --paired --length 35 --gzip --cores 2. Clean reads were aligned to human reference genome (GRCh38) combining spike-in sequences using bwa mem (version 0.7.17-r1188) with the following parameters -I 500,120,1000,20. Reads with MAPQ <1 were excluded from further analysis. Picard MarkDuplicates (version 2.18.29-SNAPSHOT) was used to identify duplicate reads. MethylDackel extract (version 0.5.0) was used for methylation calling using the following parameters -q 10 -p 13 -t 4 --mergeContext --OT 10,140,75,75 --OB 10,140,75,75. CpG sites overlapped with common SNP (dbSNP153), blacklisted regions (Amemiya, Kundaje and Boyle, 2019) and centromeres were excluded for further analysis (Rosenbloom *et al.*, 2015).

5.7.10 Analysis of cfDNA methylation by TAPS

To calculate methylation per CpG site number of modified cytosines was divided by number of all cytosines at particular position. CpG sites covered by less than 3 reads were excluded from analysis.

Number of modified and unmodified cytosines in genomic regions or genomic bins were calculated by bedtools map (Quinlan and Hall, 2010). Methylation in genomic windows or features was calculated as number of all methylated cytosines in the region divided by total number of cytosines in the region.

Genomic annotations were downloaded from ensembl database (ftp://ftp.ensembl.org/pub/release-99/gtf/homo_sapiens/Homo_sapiens.GRCh38.99.gtf.gz, ftp://ftp.ensembl.org/pub/release-97/regulation/homo_sapiens/homo_sapiens.GRCh38.Regulatory_Build.regulatory_features.20190329.gff.gz), UCSC (Rosenbloom *et al.*, 2015), RepeatMasker (<http://www.repeatmasker.org/genomes/hg38/RepeatMasker-rm405-db20140131/hg38.fa.align.gz>), Genecards (https://www.genecards.org/GeneHancer_version_4-4) and biosciencedbc (ftp://ftp.biosciencedbc.jp/archive/fantom5/datafiles/reprocessed/hg38_latest/extra/enhancer/F5.hg38.enhancers.bed.gz).

5.7.11 Differential methylation analysis

Differentially methylated CpG sites were identified by bsseq R package (Hansen, Langmead and Irizarry, 2012). Methylation was smoothed using minimum number of methylation loci in a smoothing window $n = 15$ and the minimum smoothing window $h = 100$ bp. Differentially methylated CpG sites were identified by t-test. Sites covered by less than 3 reads or in less than half of samples were excluded from analysis. Heatmaps of top differentially methylated sites were generated in R.

Number of DMC overlapping with genomic features were calculated in R and genomic enrichment was tested by Fisher test. GO analysis was performed by Enrichr (Chen *et al.*,

2013; Kuleshov *et al.*, 2016). Overlaps between PDAC DMC and DMC from RRBS public data (Shen *et al.*, 2018) were calculated in R. Statistical significance was tested using permutation test (1000 permutations, two-sided, P values computed using standard normal distribution).

5.7.12 Fragmentation analysis

The length of the DNA fragments was obtained from alignment files using Samtools (Li *et al.*, 2009). Fragmentation profiles were calculated as fraction of cfDNA fragments at 10 bp length range bins. PCA analysis and plots were generated in R.

5.7.13 Model building

All models were trained and evaluated based on leave-on-out approach. Briefly, one sample was left as a testing set and remaining were used as a training set.

Differentially methylated regions (promoters for PDAC and enhancers for HCC) were identified in the training set by t-test ($P < 0.05$, methylation difference > 0.05). Top 200 DMRs selected based on P value were selected for model training.

Differentially methylated CpG sites were identified by bsseq package (Hansen, Langmead and Irizarry, 2012) as described above. Top 1000 the most significant DMCs selected based on t-stat value and with methylation difference ≥ 0.1 were selected for model training.

Prediction model was built on selected DMCs or DMRs using cvglm, GLMnet, svm and random forest and validated on the test sample. Procedure was repeated N times, where N = number of samples.

For fragmentation-based prediction, proportion of long fragments (310-500 bp) in 10 bp length range bins. Models were built and trained by leave-one-out approach as described above using GLMnet, svm, cvglm and random forest methods.

ROC curves were prepared in R based on validation prediction scores from cvglm models for methylation-based prediction and GLMnet models for fragment length-based prediction validation.

*All sorts of things can happen when you're open to new ideas
and playing around with things.*

Stephanie Kwolek

6 Endonuclease-enrichment TAPS

6.1 Introduction

Whole genome methylation sequencing is the most comprehensive but also relatively expensive way to study DNA methylation (Plongthongkum, Diep and Zhang, 2014). In fact, as only around 4% of all cytosines in the genome are modified, even up to 65-80% of all whole genome sequencing reads do not provide any methylation information (Ziller *et al.*, 2013; Breiling and Lyko, 2015). Enrichment-based approaches implement additional steps to capture the most informative regions prior methylation sequencing and allow for cost-effective analysis of epigenetic marks in DNA (Raiber *et al.*, 2017).

DNA modifications can be enriched in various ways. The most common approaches involve enrichment by restriction digestion (Meissner *et al.*, 2005; Wang *et al.*, 2013) or by affinity binding e.g. using methylation-specific antibodies or chemical labelling (Song *et al.*, 2011; Taiwo *et al.*, 2012). Although, affinity enrichment methods allow for robust and efficient analysis of the epigenome covering over 80% of all CpG sites, their resolution is limited to around 150 bp and they do not quantify methylation at single CpG level (Maunakea *et al.*, 2010; Robinson *et al.*, 2010; Taiwo *et al.*, 2012; Thomson *et al.*, 2013). Therefore, other approaches that combine cost-effectiveness of enrichment methods but still provide information about methylation status of single base resolution have been developed.

Such an approach is Reduced Representation Bisulfite Sequencing (RRBS). RRBS involves restriction digestion of DNA to enrich regions with high CpG-density followed

by bisulfite treatment (Meissner *et al.*, 2005). Although, RRBS allows to analyse methylome at lower costs, it is restricted only to a small proportion of CpG sites located at specific sequence contexts and thus leaves a lot of methylated sites uncharacterized (Plongthongkum, Diep and Zhang, 2014). Therefore, new enrichment-based strategies to analyse DNA methylation are still needed.

Development of whole-genome sequencing TAPS (wgTAPS) opened new opportunities to design novel, improved enrichment-based methods. TAPS can be easily combined with a reduced representation approach to develop reduced representation TAPS (rrTAPS) analogous to RRBS. Moreover, products of TAPS conversion can be directly enriched by enzymatic digestion as some endonucleases are known to digest DHU or similar nucleotides (Venkhataraman *et al.*, 2001; Redrejo-Rodríguez *et al.*, 2016). These endonucleases could be used for development of endonuclease enrichment TAPS (eeTAPS) where methylated positions are directly captured by digestion of TAPS-converted DNA. Digested and size-selected fragments could be then built into a sequencing library with methylated positions enriched at the ends of sequencing reads.

EeTAPS would enable to enrich methylated CpG sites (mCpGs) in the genome independently on their sequence context and thus overcome some limitations of traditional reduced representation approaches while still provide base-resolution methylation.

6.2 Objectives

The main objectives of this project were to develop cost-effective, enrichment-based TAPS-derived method, endonuclease-enrichment TAPS (eeTAPS), and validate its performance in mESC DNA methylation sequencing in comparison to more conservative

approaches: whole-genome methylation sequencing by TAPS (wgTAPS) and reduced representation methylation sequencing by TAPS (rrTAPS).

6.3 DHU is cleaved by restriction nucleases

To develop eeTAPS we had to first identify an endonuclease able to efficient and specific cleavage of DHU in TAPS-converted DNA. We selected 10 commercially available enzymes that cleave DHU or structurally similar bases such as uracil, 5-hydroxyuracil or dihydrothymine and tested their ability to digest TAPS-converted DNA (Jiang *et al.*, 1997; Yao and Kow, 1997; Parikh *et al.*, 2000; Venkhataraman *et al.*, 2001; Ishchenko *et al.*, 2003; Gros *et al.*, 2004; Redrejo-Rodríguez *et al.*, 2016). This screening revealed that some enzymes including Uracil-specific excision reagent (USER), Endonuclease VIII, Endonuclease III and Formamidopyrimidine DNA Glycosylase (Fpg) can indeed cleave TAPS-converted DNA at various efficiencies. In contrast, Apurinic/aprimidinic Endonuclease 1 (APE1) and Uracil DNA glycosylase (UDG) left converted DNA uncleaved (Figure 6-1).

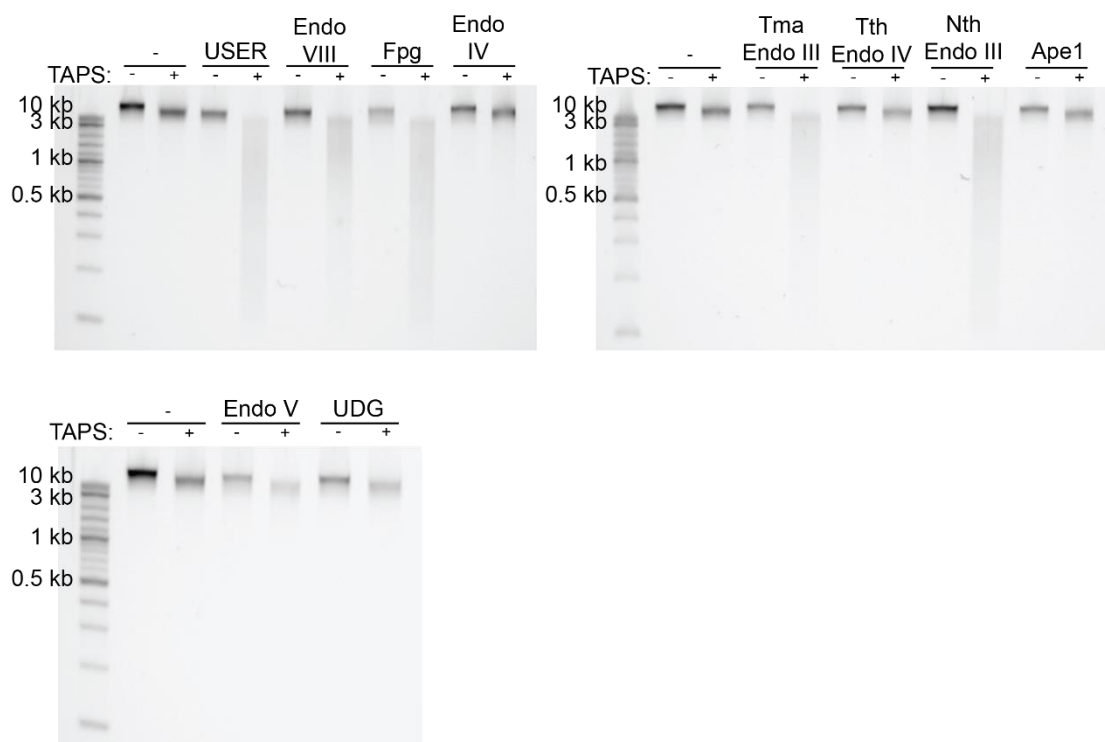


Figure 6-1. Screening for DHU-digesting endonucleases. mESC gDNA was digested before and after TAPS conversion. In TAPS-treated DNA, all 5mC are converted to DHU that can be substrate for some endonucleases. Endo VIII - endonuclease VIII, Endo IV - endonuclease IV, Tma Endo III - endonuclease III, Tth Endo IV - endonuclease IV; Nth Endo III - Nth endonuclease III; Endo V - Endonuclease V, UDG - Uracil DNA glycosylase, Ape1 - Apurinic/aprimidinic Endonuclease 1, USER - Uracil-specific excision reagent, Fpg - Formamidopyrimidine DNA Glycosylase.

We selected USER enzyme for eeTAPS development as it is commonly used, cheap and easily accessible molecular biology reagent and it cleaved DHU robustly with minimal unspecific digestion of unconverted DNA in our fragmentation assay. Matrix-assisted laser desorption/ionization mass spectroscopy (MALDI) analysis of 17 bp model DNA

oligonucleotide confirmed that DHU is a direct substrate for USER digestion in double-stranded DNA (Figure 6-2).

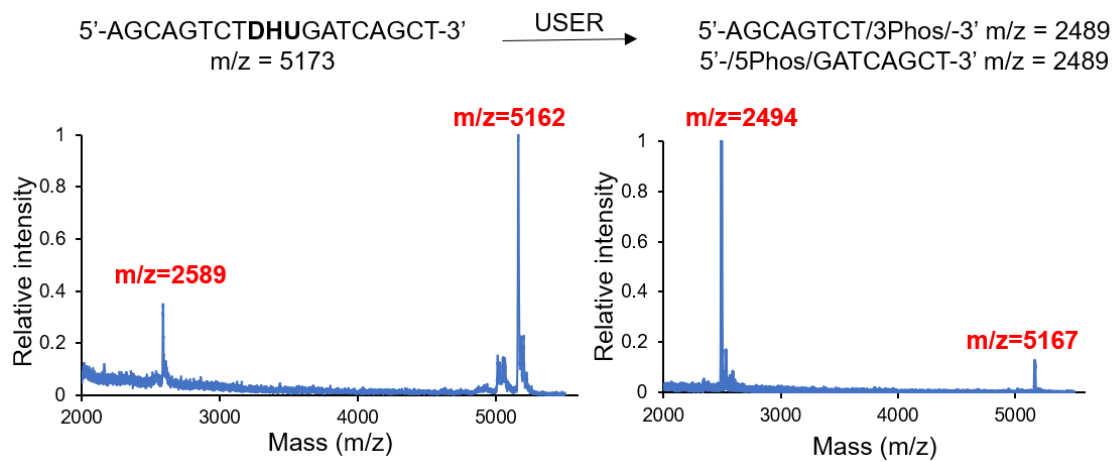


Figure 6-2. MALDI characterization of DHU-containing 17mer double-stranded model DNA before and after DHU digestion by USER enzyme. Experiment was performed once. Calculated mass was shown in black, and observed mass was shown in red.

6.4 TAPS method can be adopted to reduce-representation and endonuclease enrichment protocols

We modified TAPS to develop new, enrichment-based methods for methylome sequencing, eeTAPS and rrTAPS, and applied both methods for methylation sequencing in mESC gDNA. Since reduced representation approaches are commonly used for methylation sequencing in combination with BS-seq, we used rrTAPS as an additional control to assess overall performance and methylation enrichment by eeTAPS.

Protocol for rrTAPS was developed and implemented by Dr Anandhakumar Chandran based on the analogous method existing for bisulfite sequencing (Meissner *et al.*, 2005). Briefly, mESC gDNA was first digested with methylation-insensitive restriction enzyme, MspI, which recognizes and cleaves all CCGG sites in the genome. CpG sites are unevenly distributed in DNA and are underrepresented in most of the genome but enriched in specific regions such as CGIs (Lander *et al.*, 2001). Enzymatic digestion step together with the final size selection enables to capture regions with high CpG density

and in this way reduce sequencing costs. Cleaved DNA was then used for construction of sequencing library and underwent TAPS treatment. Final library after amplification was size-selected and sequenced on Illumina platform (Figure 6-3).

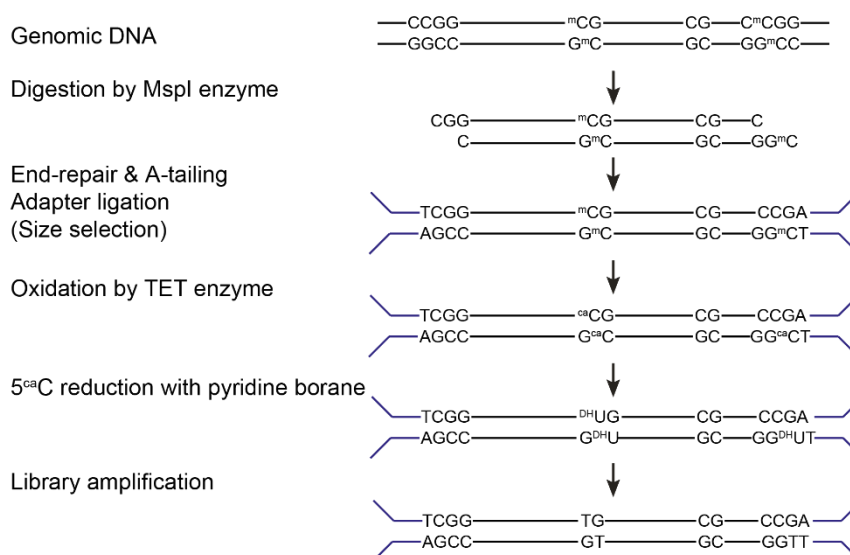


Figure 6-3. Overview of rrTAPS protocol. rrTAPS protocol was developed and implemented by Dr Anandhakumar Chandran.

Another approach, eeTAPS, was developed by combining TAPS with endonuclease enrichment. For eeTAPS preparation, unfragmented mESC gDNA was subjected to mTet1CD oxidation and pyridine borane reduction and subsequently digested by USER enzyme. USER cleavage produced fragments with broad length distribution from 100 bp to 10 kb with short fragments likely corresponding to densely methylated parts of the genome and long fragments corresponding to sparsely methylated regions. Afterward, digested fragments ranging from 200 bp to 1 kb were size selected to enrich for moderately modified DNA regions and constructed into next generation sequencing library. Upon this treatment, modified cytosines were expected to be enriched at the ends of sequencing reads. Modification values for individual cytosines were calculated as a ratio between reads cleaved at the particular position and sum of reads cleaved and covering this site (Figure 6-4).

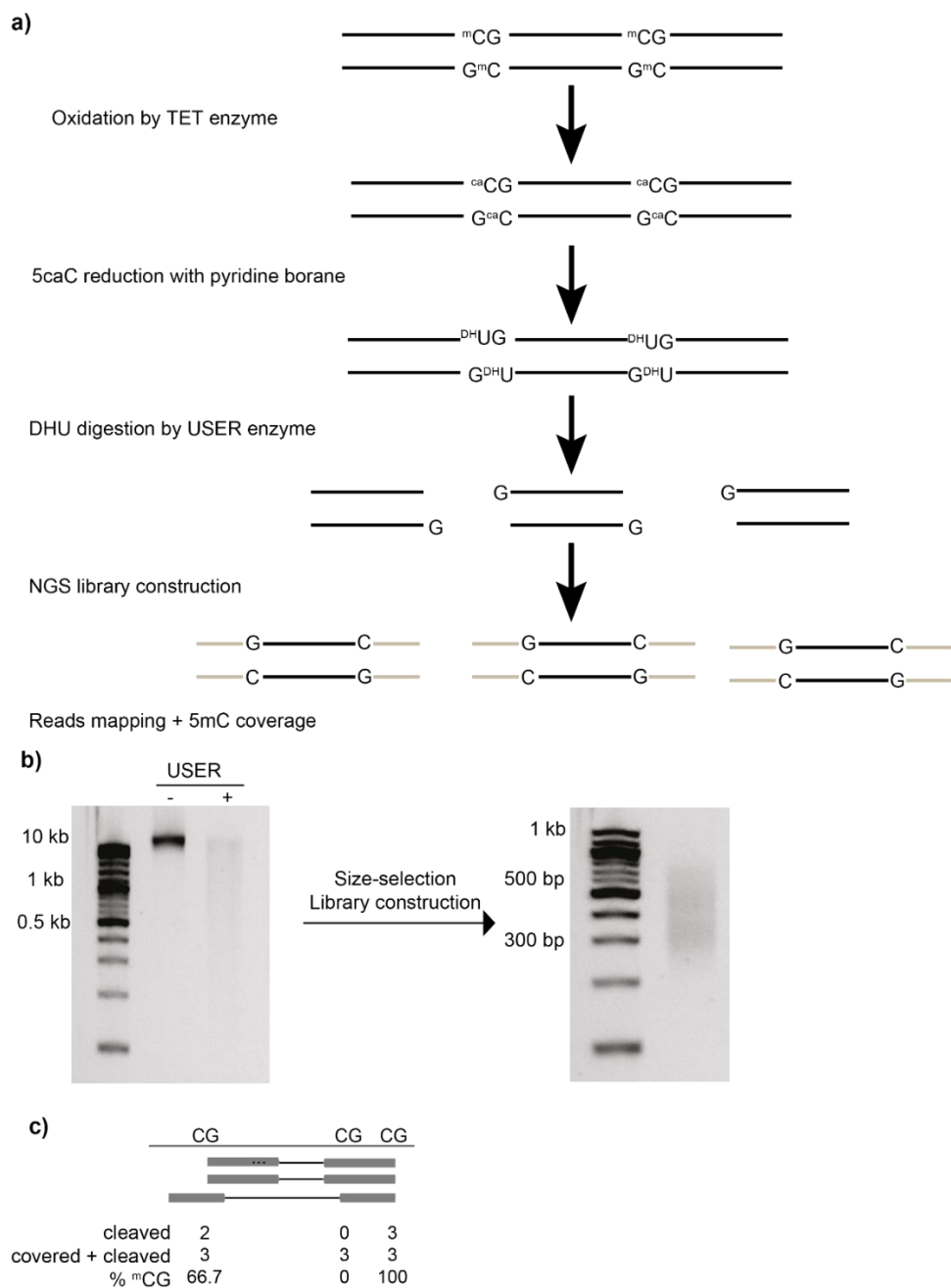


Figure 6-4. (a) Overview of eeTAPS method. (b) Representative image of TAPS-treated mESC gDNA without and with digestion and image of final size selected eeTAPS sequencing library. (c) The principle of CpG methylation quantification for eeTAPS.

6.5 eeTAPS provides reliable methylation information

To test whether eeTAPS provides accurate and reliable methylation estimates, we tested its performance on 4 kb model DNA with CpG methylation at CCGG context and

compared it with the results of bisulfite sequencing. We confirmed that methylation values measured by both methods were at perfect agreement (Pearson correlation coefficient $r = 0.98$) (Figure 6-5, a-b). Moreover, USER enzyme specifically enriched for methylated sites as no significant methylation was detected in control, unconverted DNA digested by the enzyme (Figure 6-5, b). These results proved that eeTAPS provides accurate and specific methylation information.

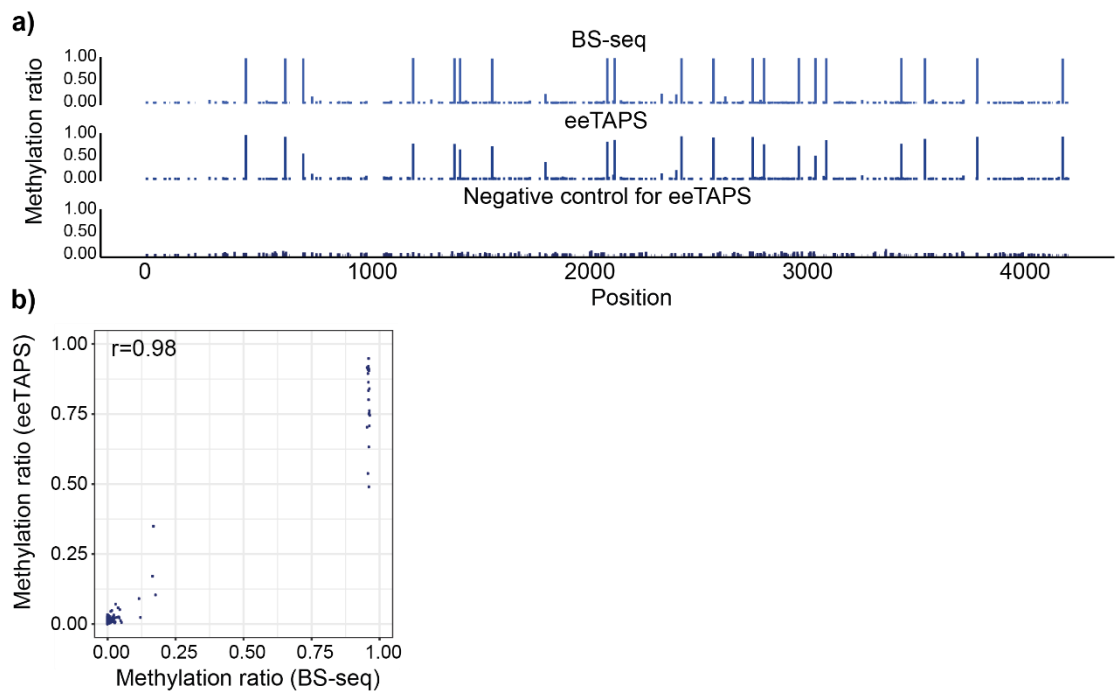


Figure 6-5. Evaluation of eeTAPS method. (a) Comparison of methylation levels provided by eeTAPS and bisulfite sequencing on model 4 kb DNA. Negative control for eeTAPS is library constructed on unconverted DNA cleaved by USER enzyme. (b) Correlation plot of per CpG methylation levels estimated by eeTAPS and bisulfite sequencing on model 4 kb. Data analysis was performed based on scripts developed by Dr Jingfei Cheng.

Having validated eeTAPS method on model DNA, we applied this protocol to profile methylation in mESC gDNA and compared it with wgTAPS and rrTAPS. We confirmed that eeTAPS enriched for methylated CpG sites (mCpGs), as nearly 85% of sequencing reads generated by this method ended with C or G (Figure 6-6, a). Additional analysis revealed that over 70% of cleaved sites overlapped with CpG sites (Figure 6-6, b).

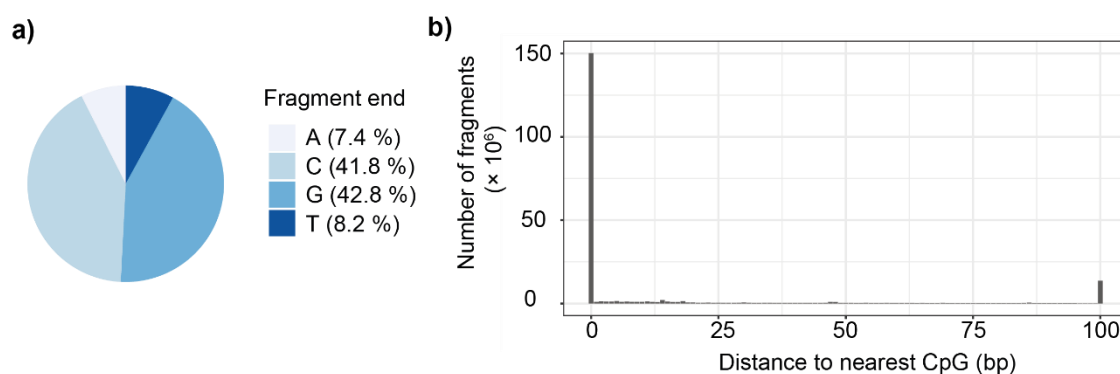


Figure 6-6. eeTAPS cleavage summary. (a) Frequency of nucleotides at the ends of eeTAPS sequencing reads. (b) Distribution of distance between cleavage events and the nearest CpG site. Data analysis was performed based on scripts developed by Dr Jingfei Cheng.

Subsequently, we compared percentage of all mESC CpG sites covered by three TAPS-based methods. Not surprisingly, wgTAPS covered 92.1% (19.3M) of all CpG sites in mESC genome providing the most comprehensive methylation map. On the other hand, eeTAPS coverage was just slightly worse as this method still detected majority of CpG sites (74.2%, 15.2M). In contrary, rrTAPS covered only around 7.2% (1.5M) of mESC CpG sites which is consistent with a proportion of CpG sites covered by the analogous, bisulfite-based method – RRBS (Plongthongkum, Diep and Zhang, 2014). This together demonstrated that endonuclease enrichment of methylated sites in eeTAPS allows for detection of majority of CpG sites and covers over 10-fold more positions than traditional reduced representation approach.

We then analysed mCpGs identified by eeTAPS, rrTAPS and wgTAPS. We first confirmed that eeTAPS provides reproducible methylation as most of mCpGs in mESC were detected by two independent replicates (Figure 6-7, a). Moreover, eeTAPS detected over 80% of mCpGs detected by wgTAPS compared to around 20% detected by rrTAPS (Figure 6-7, b). There was good correlation between methylation at single-CpG level quantified by eeTAPS and wgTAPS ($r = 0.56$) and almost perfect correlation between methylation levels calculated by rrTAPS and wgTAPS ($r = 0.92$) (Figure 6-7, c-d). Worse correlation between CpG methylation measured by eeTAPS and wgTAPS could be

caused by less accurate quantification provided by the former method which may be affected by cleavage efficiency or PCR amplification bias. Nevertheless, all these results demonstrated that eeTAPS can specifically and reproducibly detect majority of CpG sites in the genome and provides semi-quantitative information about methylation level of individual cytosines.

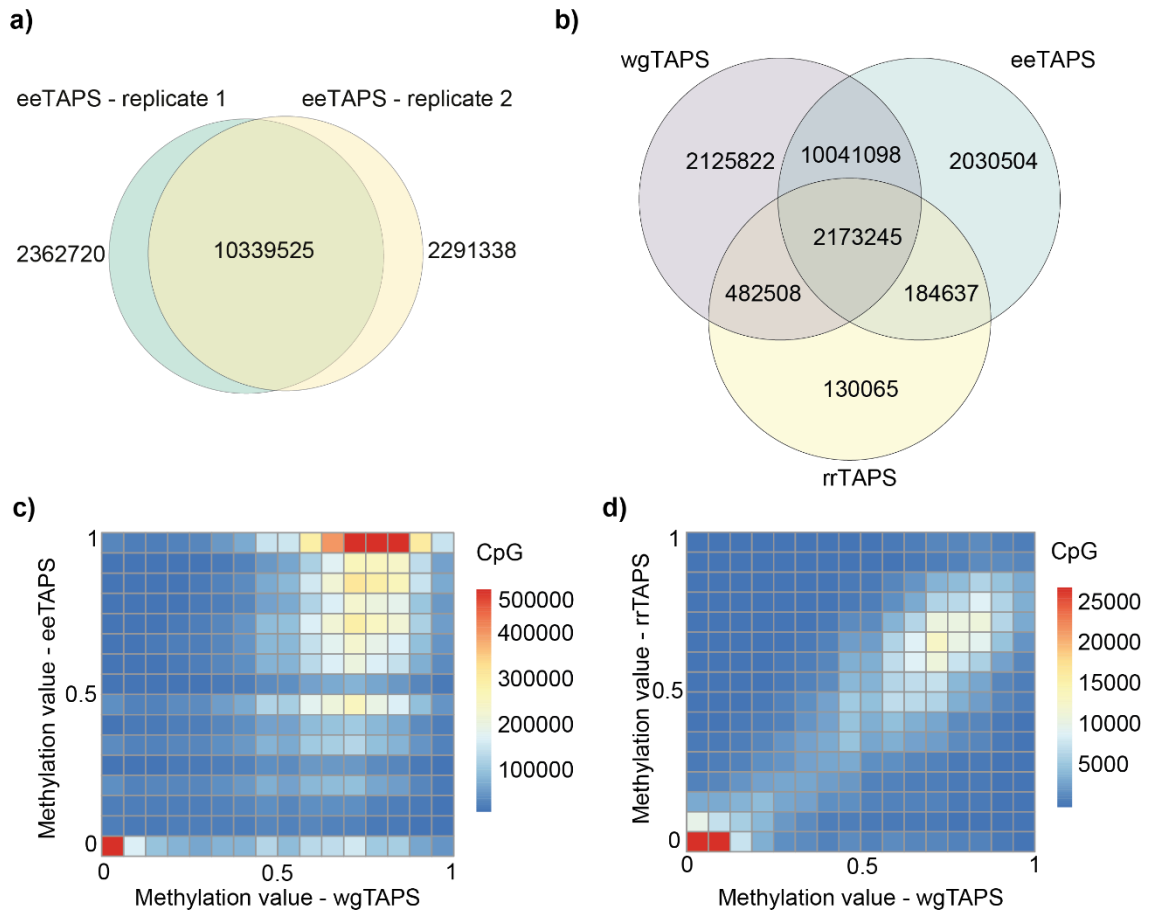


Figure 6-7. (a) Number of mCpGs detected in mESC gDNA by two independent replicates of eeTAPS. (b) Overlap of mCpGs detected by wgTAPS, eeTAPS and rrTAPS. In wgTAPS and eeTAPS CpGs with methylation levels higher than 1st quantile were defined as methylated. In rrTAPS, mCpGs were defined based on wgTAPS cut-off. In summary, CpG sites with methylation higher than 0.5 were classified as methylated in wgTAPS and rrTAPS and CpGs with methylation higher than 0.28 were classified as methylated in eeTAPS. (c) Correlation heatmap of single-CpG methylation measured by wgTAPS and eeTAPS ($r = 0.56$). CpG sites were grouped by modifications level. CpG sites covered by less than five reads were excluded from analysis. (d) Correlation heatmap of single-CpG methylation measured by wgTAPS and rrTAPS ($r = 0.92$). CpG sites were grouped by modifications level. CpG sites covered by less than five reads were excluded from analysis. Data analysis was performed based on scripts developed by Dr Jingfei Cheng.

6.6 eeTAPS provides comprehensive information about methylation distribution

We sought to compare genomic distribution of all CpG sites covered by wgTAPS, eeTAPS and rrTAPS. We found out that substantial fraction of CpG sites detected by eeTAPS and wgTAPS were located within heterochromatin (71% and 65.6%, respectively) while most of CpG sites covered by rrTAPS overlapped with promoter regions (Figure 6-8, a).

Subsequently, we analysed methylation signatures provided by eeTAPS and wgTAPS in various genomic features (Bogu *et al.*, 2016). For wgTAPS, methylation in a genomic feature was calculated as an average methylation of all CpG sites in a region. For eeTAPS, methylation level was calculated as a ratio between mCpGs and all CpG sites in a region. Both methods shown similar methylation patterns across chromosomes as well as CGIs with correlation level of 0.87 and 0.81, respectively (6-8, b-d). However, eeTAPS exhibited consistent tendency for overestimating methylation compared to wgTAPS. This may suggest that eeTAPS underestimates number of unmethylated CpG sites, possibly due to lower amplification efficiency of long, unmodified DNA fragments that remain uncleaved after USER digestion.

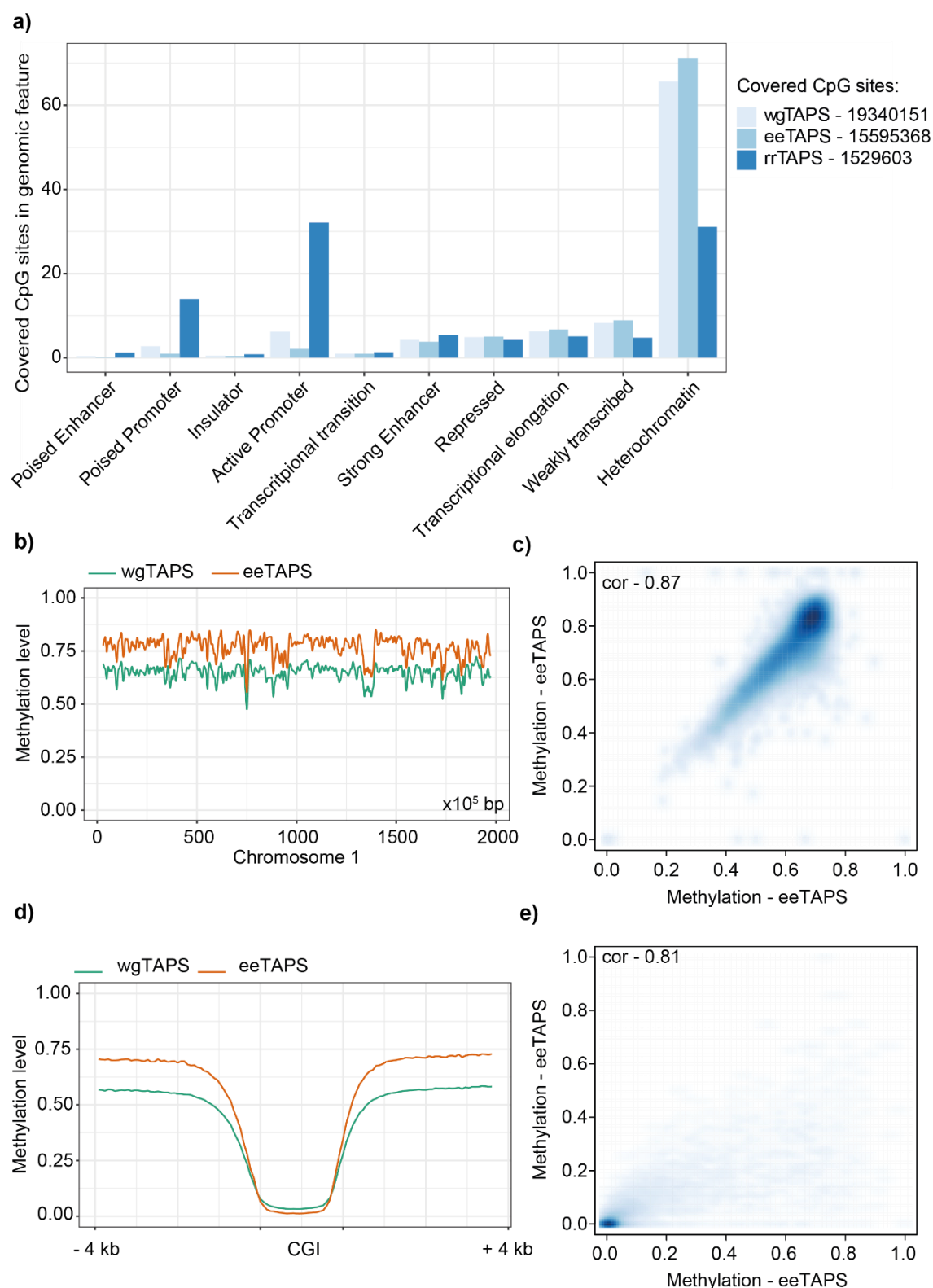


Figure 6-8. (a) Fraction of CpG sites covered by eeTAPS, wgTAPS and rrTAPS overlapping with different genomic features. Total number of covered CpG sites for each method is shown in the legend. (b) Distribution of modification levels in chromosome 1 measured by eeTAPS and wgTAPS. Moving average CpG modification value was calculated per 100 kb windows along chromosome 1 (bw = 10). (c) Correlation of methylation in 100 kb genomic bins calculated as described in (b) between eeTAPS and wgTAPS (Pearson correlation $r = 0.87$). (d) Average

modification levels in CGI and 4 kb flanking regions surrounding regions measured by eeTAPS and wgTAPS. (e) Correlation of methylation in CGI and 4 kb flanking regions between eeTAPS and wgTAPS (Pearson correlation $r = 0.81$). Data analysis was performed based on scripts developed by Dr Jingfei Cheng.

We also looked at methylation levels in different chromatin features and found that both methods can capture feature-specific methylation differences such as hypermethylation in heterochromatin and depleted methylation in active promoters (Figure 6-9)

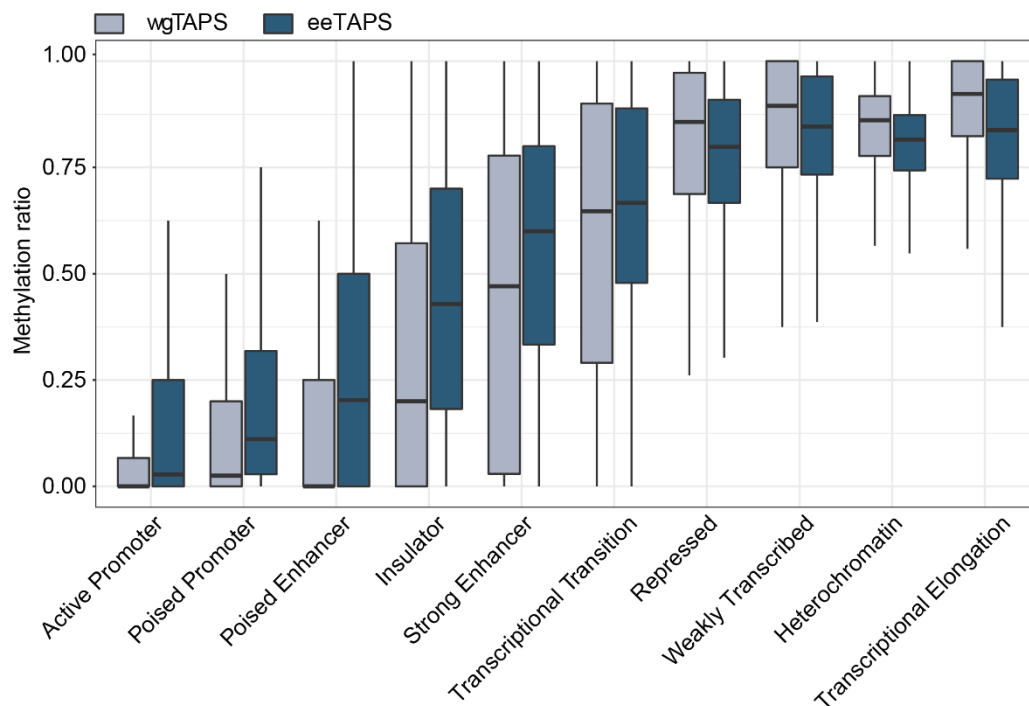


Figure 6-9. Methylation level measured by wgTAPS and eeTAPS in various genomic features. Data analysis was performed based on scripts developed by Dr Jingfei Cheng.

Repressive effect of promoter methylation on gene expression is well-described in the literature (Borgel *et al.*, 2010; Ficz *et al.*, 2011; Leung *et al.*, 2015). Therefore, we aimed to compare methylation levels measured by eeTAPS and wgTAPS with gene expression. We used public data (GSE72855 dataset) to divide genes into 4 groups depending on their expression levels and compare them with their methylation levels around transcriptional start site (TSS). Not surprisingly, we found that genes with lower methylation levels exhibited higher expression levels, while genes with higher methylation were lowly expressed (Figure 6-10).

To sum up, eeTAPS, similar to wgTAPS, accurately captured characteristic methylation patterns in various genomic features as well as in differentially expressed genes.

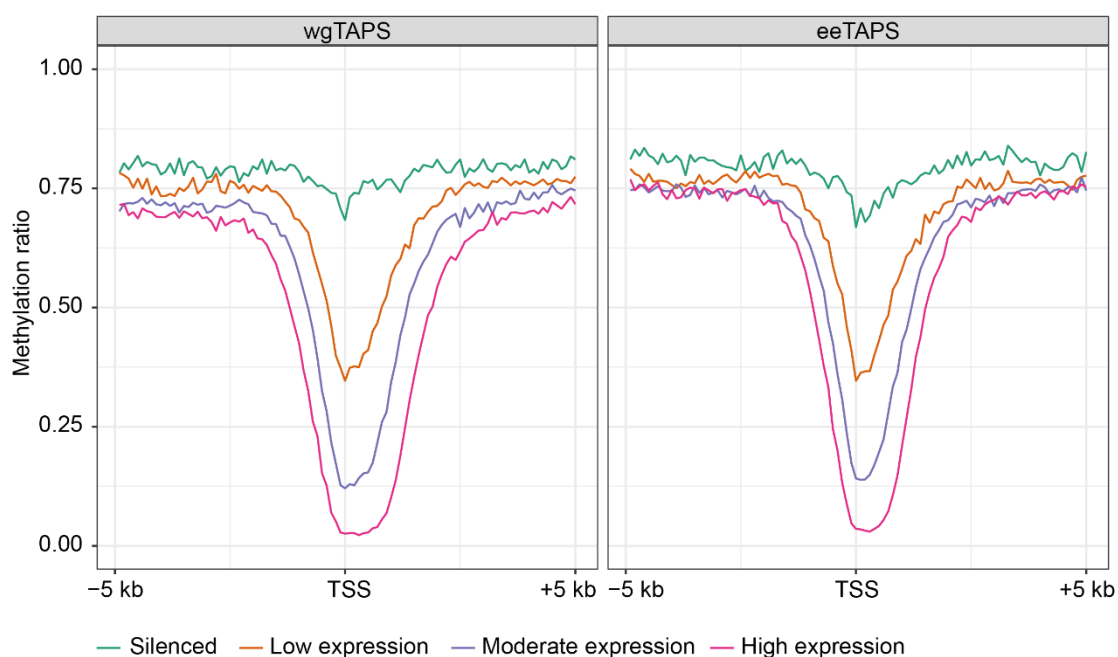


Figure 6-10. Methylation distribution around TSS measured by wgTAPS and eeTAPS in genes categorized according to their expression level (GSE72855 dataset). Silenced genes were defined as genes with Reads Per Kilobase of Transcript, per Million Mapped Reads (RPKM) equal to 0. Low expression, moderate expression and high expression genes were defined based on RPKM quantile cutoffs: low expression - RPKM > 0 ~ 1/3 quantile; moderate expression – RPKM = 1/3 ~ 2/3 quantile; high expression - RPKM = 2/3 ~ 3/3 quantile. Data analysis was performed based on scripts developed by Dr Jingfei Cheng.

6.7 eeTAPS can be successfully applied for low input applications

As TAPS conditions are mild and non-damaging, we evaluated performance of endonuclease enrichment protocol on low-input samples. We prepared eeTAPS libraries from 1 ng, 10 ng and 50 ng of gDNA and compared them with standard library prepared from 200 ng of gDNA. As we used shallow sequencing for low input samples, we subsampled standard eeTAPS library to 2X to avoid a coverage bias due to higher sequencing depth.

We were able to detect 47% of mCpGs covered by wgTAPS in eeTAPS library prepared from 50 ng and 27% when prepared from 1 ng (Figure 6-11, a). We found that low-input

libraries provided methylation patterns in 100 kb genomic bins highly similar to high-input eeTAPS with correlation 0.95, 0.92, 0.88 for 50 ng, 10 ng and 1 ng respectively (Figure 6-11, b-c).

To conclude, eeTAPS can be useful tool for cost-effective, accurate methylation analysis in low input samples.

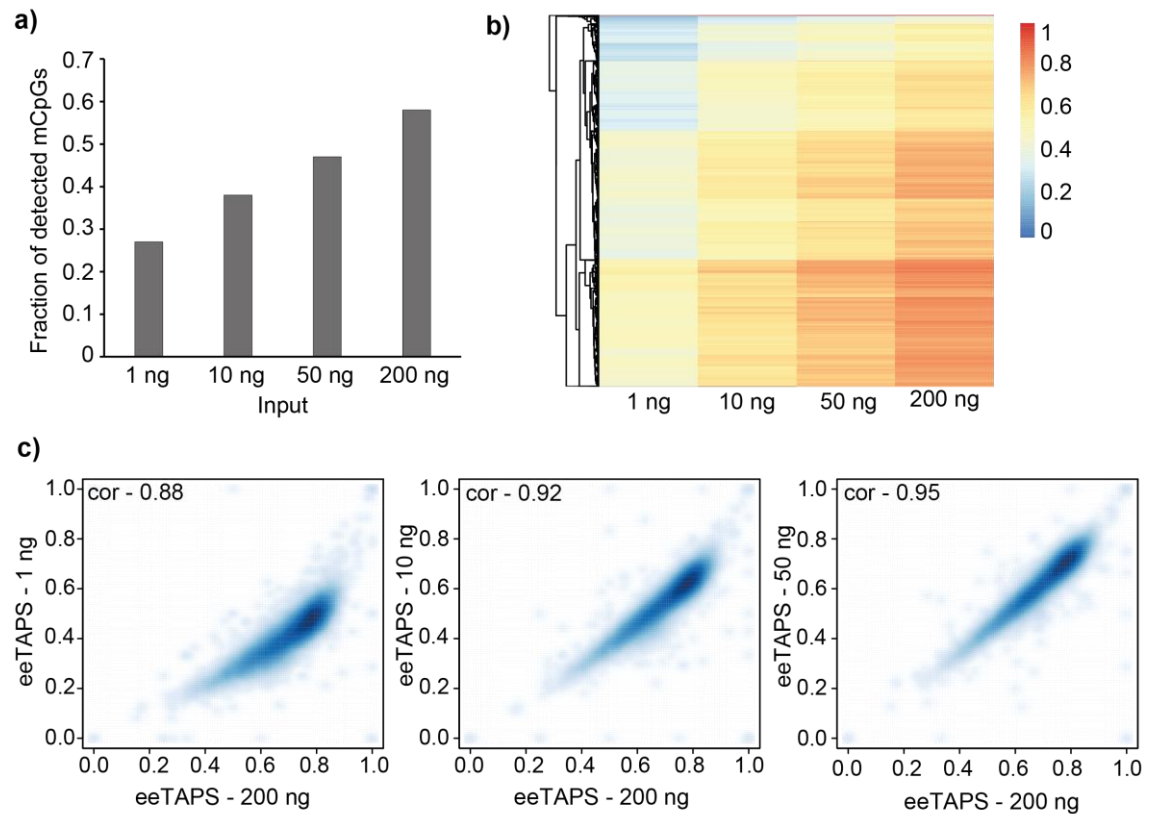


Figure 6-11. Low input eeTAPS. (a) Fraction of mCpGs identified by wgTAPS detected in eeTAPS libraries prepared from 1 ng, 10 ng, 50 ng and 200 ng of mESC gDNA. (b) Heatmap of methylation in 100 kb genomic bins measured by eeTAPS libraries prepared from different DNA input. (c) Correlation of methylation in 100 kb genomic bins calculated between standard eeTAPS and low input eeTAPS libraries. Pearson correlation ratios are shown on the plots. Data analysis was performed based on scripts developed by Dr Jingfei Cheng.

6.8 eeTAPS is cost-effective approach for methylation sequencing

Since eeTAPS was designed as cost-effective approach compared to wgTAPS, we aimed to assess what is the lowest sequencing depth required for eeTAPS. We found that with higher sequencing depth we were able to detect higher number of mCpG sites. However, even at 4X sequencing, we were still able to cover by at least 2 sequencing reads 74%

mCpGs detected by wgTAPS (Figure 6-12, a). We also demonstrated that even at the lower depth, we could still observe high methylation correlation across genomic bins and CGIs (Figure 6-12, b). This demonstrates that eeTAPS is indeed cost-effective approach that provides accurate and global methylation patterns and cover most of mCpGs even at low sequencing depth around 3-4X.

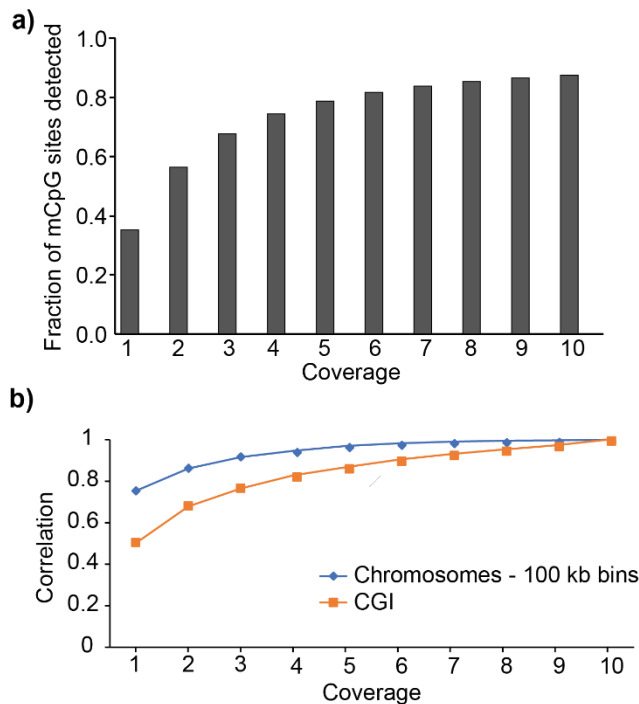


Figure 6-12. Effect of sequencing depth on eeTAPS. (a) Fraction of mCpGs identified by wgTAPS covered by eeTAPS after subsampling to different sequencing depth from 1-10X. (b) Correlation of methylation in 100 kb genomic bins and CGI measured by eeTAPS after subsampling to sequencing depth of 1-10X. Data analysis was performed based on scripts developed by Dr Jingfei Cheng.

6.9 Discussion

Although wgTAPS has been demonstrated to be reliable and accurate, bisulfite-free method for base-resolution methylation sequencing, it is also relatively expensive approach. Therefore, we sought to combine TAPS with enrichment-based approaches for cost-effective methylome sequencing.

We first combined TAPS with reduced representation protocol already known for bisulfite sequencing (Meissner *et al.*, 2005). We confirmed that this method can

accurately estimate CpG methylation with high correlation with wgTAPS ($r = 0.92$). On the other hand, similar to RRBS, rrTAPS covers less than 10% of CpG sites in the mouse genome thus methylation status of majority of positions is left uncharacterised (Meissner *et al.*, 2005; Plongthongkum, Diep and Zhang, 2014). rrTAPS enrichment is based on restriction digestion of DNA to enrich for regions with high density of CpG sites. Therefore, it is restricted to some specific parts of the genome recognizable by the restriction enzymes used for analysis such as CGI. In our rrTAPS we used single enzyme, MspI, that cleaves CpG sites in CCGG context and demonstrated that this approach enabled high enrichment of promoter regions. Potentially, combination of enzymes could help to increase number of CpG sites covered in rrTAPS to improve coverage in other genomic features. In the future a cocktail of restriction endonucleases with different restriction sites could be used to increase the number of detectable CpG sites and enable more comprehensive analysis of the genome (Martinez-Arguelles, Lee and Papadopoulos, 2014; Martin-Herranz *et al.*, 2017).

Apart from rrTAPS, we also aimed to develop completely new protocol for methylation enrichment by endonuclease enrichment. As wgTAPS converts modified cytosines directly, we envisioned that product of conversion – DHU, can be enriched by endonuclease digestion. We identified number of endonucleases with potential ability to cleave DHU (Venkhataraman *et al.*, 2001; Redrejo-Rodríguez *et al.*, 2016) and tested their digestion efficiency at TAPS-converted DNA. We selected USER enzyme based on efficient and robust fragmentation. We demonstrated that eeTAPS covers 74% of mCpGs by at least two reads at shallow, 4X sequencing depth, providing comprehensive methylation picture at low sequencing costs. As eeTAPS is based on milder, non-damaging conditions, we successfully applied it to low input samples and demonstrated

that eeTAPS can accurately detect methylation in as low as 1 ng of input DNA at low sequencing costs.

eeTAPS methylation in genomic bins and genomic regions correlated well with quantification based on wgTAPS, which proves its ability to accurately quantify methylation patterns. On the other hand, methylation correlation between eeTAPS and wgTAPS was significantly worse at single-CpG level and was lower than for rrTAPS ($r = 0.56$ vs. $r = 0.92$). We also found that eeTAPS has a tendency to overestimate methylation compared to wgTAPS. On the other hand, even with the current conditions eeTAPS could still accurately capture characteristic methylation patterns in various genomic regions such as high methylation levels in heterochromatin, hypomethylation of CGIs or distinct promoter methylation states in differentially expressed genes. This suggests that although methylation quantification by eeTAPS may be less accurate compared to wgTAPS or rrTAPS, this method still provides reliable information about genomic modification levels.

In our eeTAPS protocol we selected for 200-1000 bp fragments after USER digestion as we suspected that they represent moderately methylated regions. However, this range may not be suitable for all experimental settings. In the future size selection can be expanded to other fragment ranges to select for more specific enrichment. For instance, shorter fragments could be selected to restrict analysis only to highly methylated regions or longer fragments to restrict sequencing to sparsely methylated regions in the genome, similarly to enhanced RRBS (Akalin *et al.*, 2012).

Although we proved that eeTAPS can be a powerful cost-effective tool to measure DNA methylation it is important to note, that eeTAPS is a semi-quantitative method and may be affected by various factors. As eeTAPS methylation at CpG sites is defined as number of cleavage events divided by number of total reads covering the position, unspecific

cleavage events or underestimated number of unmodified cytosines may affect quantification. Given that we observed higher methylation values in eeTAPS, it is possible that unmethylated cytosines are underrepresented. Unmethylated sites remain unconverted by TAPS and are not cleaved by USER, therefore they may be associated with longer fragment length and can be removed by applied size selection cut-off or amplified with less efficiency. This issue could be addressed in the future developments and various size selection cut-offs as well as optimised amplification protocols may be tested to improve accuracy of methylation quantification by eeTAPS.

Another possible issue is USER efficiency and specificity at cleaving DHU. We selected USER enzymes based on TAPS-converted fragmentation results. Based on our assessment we concluded that USER enzyme cleaves converted DNA with the best efficiency generating the shortest fragments in the lower length range compared to other enzymes tested. On the other hand, we have not accurately quantified USER efficiency on DHU cleavage. This could be tested in the future with more quantitative assays. For instance, USER digestion efficiency could be potentially quantified by digestion of model DNA sequence containing DHU and quantification of digested and undigested bands on the gel or in more precise ways by qPCR. Also, we have not compared USER digestion at various sequence contexts so we cannot exclude potential context-dependent cleavage bias caused by USER sequence preference. In the future, more comprehensive analysis and optimisation of restriction endonucleases may be performed to better characterise and optimise USER-mediated DHU enrichment or to identify endonucleases that outperform USER enzyme.

6.10 Concluding remarks

We demonstrated that previously developed TAPS method can be combined with enrichment-based approaches for cost-effective methylation sequencing. We proposed eeTAPS as a powerful and promising tool for 5mC enrichment and compared it with wgTAPS and rrTAPS. Our results suggest that eeTAPS can accurately characterise methylation status of majority of CpG sites in the mouse genome at lower sequencing costs. On the other hand, as most of methods, eeTAPS has some limitations including semi-quantitative character and overestimation of methylation values. Those issues may be addressed in the future developments to improve eeTAPS performance and enable its broader application.

6.11 Materials and Methods

6.11.1 Preparation of spike-in controls

A 4 kb spike-in control was prepared by PCR amplification of the pNIC28-Bsa4 plasmid (Addgene, cat. no. 26103) in a reaction containing 1 ng DNA template, 0.5 μ M primers and 1X Phusion High-Fidelity PCR Master Mix with HF Buffer (Thermo Scientific). Primer sequences are listed in Table 6-1. The PCR product was purified by Zymo-IC column (Zymo Research) and methylated by HpaII Methyltransferase (New England Biolabs) for 2 h at 37°C in a 50 μ L reaction. Methylated product was purified with 1X AMPure XP beads (Beckman Coulter) according to the manufacturer's protocol. Fully CpG-methylated λ -DNA was prepared by methylation of unmethylated λ -DNA (Promega) with M. SssI enzyme (New England Biolabs) as described in section 4.11.2.

6.11.2 Preparation of carrier DNA

Carrier DNA was prepared by PCR amplification of the pNIC28-Bsa4 plasmid (Addgene, cat. no. 26103) in a reaction containing 1 ng DNA template, 0.5 μ M primers and 1X Phusion High-Fidelity PCR Master Mix with HF Buffer (Thermo Scientific). Primer sequences are listed in Table 6-1. The PCR product was purified by Zymo-IC column (Zymo Research), fragmented by Covaris M220 and purified on 0.9 $\times$ AMPure XP beads to select for 200-500 bp fragments.

Table 6-1. Primer sequences used to amplify 4 kb spike-in model DNA and carrier DNA.

Template	Primer	Sequence (5' to 3')
4 kb model DNA	4 kb-F	CATCGAGCATCAAATGAAACTGC
	4 kb-R	ACGTTATACGATGTCGCAGAGT
Carrier 2 kb	2 kb-F	AGGCAACTTTATGCCCATGCAA
	2 kb-R	CCAAGGGGTTATGCTAGTTATTGC

6.11.3 Expression and purification of mTet1CD

The expression and purification of mTet1 catalytic domain (mTet1CD) were done as described in section 4.11.5.

6.11.4 mESCs culture and isolation of genomic DNA

The mESCs culture and isolation of genomic DNA were done as described in section 4.11.12.

6.11.5 mTet1CD oxidation

The mTet1CD oxidation was done as described in section 4.11.8.

6.11.6 eeTAPS

mESC genomic DNA (200 ng, 50 ng, 10 ng or 1 ng) was spiked with 0.05% 4 kb control methylated in CCGG sequence context and oxidised by mTet1CD as described above.

Subsequently, oxidized DNA samples in 35 μ l of water were reduced in a 50 μ l reaction containing 600 mM sodium acetate solution (pH 4.3) and 1 M pyridine borane for 16 h at 37°C and 850 r.p.m. in an Eppendorf ThermoMixer. The product was purified using Zymo-Spin columns. Converted samples were digested in a 20 μ L reaction containing 2 U of USER enzyme (New England Biolabs) in CutSmart buffer for 1 h at 37°C and size-selected on 0.35-1 $\times$ AMPure XP beads. End-repair and A-tailing reactions, and ligation of Illumina Multiplexing adapters were prepared with KAPA HyperPlus kit according to the manufacturer's protocol. To prepare the control library, 200 ng of unconverted mESC gDNA with spike-in controls was digested by USER enzyme, size-selected and used for library construction as described above. The final sequencing libraries were amplified with KAPA HiFi HotStart ReadyMix for 6 cycles (for 200 ng input), 8 cycles (50 ng input), 10 cycles (10 ng input) or 14 cycles (1 ng input) and size-selected on 0.35-1 $\times$ AMPure XP beads. Final libraries were paired-end 80 bp sequenced on a NextSeq 500 sequencer (Illumina) together with other sequencing libraries. Sequencing statistics are shown in Appendix 6 Supplementary Table 6-1.

6.11.7 rrTAPS

1 μ g mESC gDNA was spiked with 1% CpG-methylated lambda and digested by Fast digest Msp1 enzyme (Thermo Scientific) in 50 μ L reaction for 30 min at 37°C. Digested DNA was purified by the phenol/chloroform precipitation method. End-repair and A-tailing reactions, and ligation of Illumina Multiplexing adapters were prepared with NEBNext® Ultra™ II DNA Library Prep Kit according to the manufacturer's protocol. The ligated library was then purified on 1.6 $\times$ AMPure XP beads and run on a 1% agarose gel. DNA fragments from 100-400 bp were excised and purified by Monarch® DNA Gel Extraction Kit following the manufacturer's protocol. The adapter-ligated sample was

spiked with 100 ng of carrier DNA and double oxidised by mTet1CD as described above. Oxidized DNA in 35 µl of water was reduced in a 50 µl reaction containing 600 mM sodium acetate solution (pH 4.3) and 1 M pyridine borane for 16 h at 37°C and 850 r.p.m. in an Eppendorf ThermoMixer. The product was purified using Zymo-Spin columns. The final sequencing library was amplified with KAPA HiFi Uracil (+) Master Mix for 6 cycles and purified on 1× AMPure XP beads. Final libraries were paired-end 80 bp sequenced on a NextSeq 500 sequencer (Illumina) together with other sequencing libraries in Appendix 6 Supplementary Table 6-2.

6.11.8 Screening for DHU digesting endonucleases

1 µg mESC gDNA was enzymatically oxidised by mTet1CD as described before. Subsequently, oxidized DNA in 35 µl of water was reduced in a 50 µl reaction containing 600 mM sodium acetate solution (pH 4.3) and 1 M pyridine borane for 16 h at 37°C and 850 r.p.m. in an Eppendorf ThermoMixer. The product was purified using Zymo-Spin columns. 40 ng of TAPS converted or unconverted DNA were then digested by the following enzymes according to the manufacturer's protocols (all from New England Biolabs): USER (Cat. No. M5505S), Endonuclease IV (Cat. No. M0304S), Tma Endonuclease III (Cat. No. M0291S), Endonuclease V (Cat. No. M0305S), UDG (Cat. No. M0280S), Tth Endonuclease IV (Cat. No. M0294S), Fpg (Cat. No. M0240S), Endonuclease III (Nth) (Cat. No. M0268S), Endonuclease VIII (Cat. No. M0299S), APE1 (Cat. No. M0282S). Digestion products were purified on 1.8× AMPure XP beads following the manufacturer's instructions and 10 ng of each product were run on a 2% agarose gel.

6.11.9 Data analysis for eeTAPS

Raw sequenced reads were processed with TrimGalore (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) to perform adapter and quality trimming with the following parameters: --paired --length 35. Cleaned reads were aligned using bwa mem 0.7.17-r1188 (Li and Durbin, 2010) with default parameters. For the 4 kb model DNA, the pNIC28-Bsa4 sequence from 2627 to 6911 was used as reference. For mESC gDNA, the mm9 genome was used as reference. Only properly mapped read pairs (Read 1 with flag assigned as 83 or 99) were extracted to compute coverage with bedtools v2.27.1 (Quinlan and Hall, 2010) for both endpoints and read-through of the whole fragments, and un-cleaved sites were also taken into consideration when calculating the cleavage fraction. The detailed computational pipeline to analyse eeTAPS can be found here <https://gitlab.com/jfeicheng/userenrich>. Two technique replicates were sequenced for eeTAPS. When analysing the effect of sequence depth on eeTAPS, the alignment files from two replicates were merged and then sub-sampled by fraction from 0.1 to 1 with samtools view (Li *et al.*, 2009).

6.11.10 Data analysis for rrTAPS

Raw sequenced reads were processed with seqtk (<https://github.com/lh3/seqtk>) trimfq -b 2 to trim 2 bp from the left of each read. Astair 3.2.7 (<https://pypi.org/project/asTair/>) was used to process rrTAPS. Cleaned reads were aligned using astair align with mm9 genome as reference. Methylated CpGs were extracted with astair call.

6.11.11 Comparison of wgTAPS, eeTAPS and rrTAPS in mESC

wgTAPS data were downloaded from GSE112520. Only CpG sites covered with at least 4 reads were considered as covered CpG sites. The number of methylated CpG sites was

defined according to the following criteria: CpG methylation level > 1st quartile of all CpG methylation level (0.5 for wgTAPS and 0.28 for eeTAPS). The genome was divided into non-overlapping 100kb windows with bedtools. The CpG island track was downloaded from <http://hgdownload.soe.ucsc.edu/goldenPath/mm9/database/cpgIslandExt.txt.gz>. The gene annotation file was downloaded from <http://hgdownload.soe.ucsc.edu/goldenPath/mm9/database/refGene.txt.gz>. For wgTAPS, the average methylation was used to assign methylation in each window. For eeTAPS, CpG sites with cleavage fraction higher than 0.28 were designated as methylated, while sites below this cut off were designated as unmethylated, and the methylation level for each bin was thus measured as the # methylated CpG / (# methylated CpG + # unmethylated CpG). Expression data from the e14 mESC cell line was taken from GEO entry GSE72855 (Neri *et al.*, 2017) and used to categorize genes into four groups according to their expression levels.

*I write entirely to find out what I'm thinking,
what I'm looking at,
what I see and what it means.
What I want and what I fear.*

Joan Didion

7 Discussion

In the final chapter I would like to discuss the most important findings of my research project in a broad context of epigenetics and early cancer detection. In particular I will emphasise our contribution to cfDNA studies, indicate future research directions and identify potential limitations.

Early cancer detection is one of the greatest challenges of science and medicine today. Liquid biopsy assays focused on cfDNA analysis are considered promising approaches for development of novel diagnostics tools (Wan *et al.*, 2017). Current studies have demonstrated that various characteristics of cfDNA including mutations (Cohen *et al.*, 2018), methylation (Shen *et al.*, 2018; Chen *et al.*, 2020; M. C. Liu *et al.*, 2020) and fragmentation patterns (Mouliere *et al.*, 2018; Cristiano *et al.*, 2019) can be used for cancer prediction. Some studies have already demonstrated that cfDNA methylation assays may overcome limitations of genetic analysis having better sensitivity at early stage and providing information of possible tumour location (Lehmann-Werman *et al.*, 2016; S. Guo *et al.*, 2017; Moss *et al.*, 2018; Oxnard *et al.*, 2019).

Unfortunately, cfDNA epigenetics is still not well understood. Most of current methylation studies were focused on targeted analysis of cancer-specific genes (Chen *et al.*, 2020) or used bisulfite-based technologies including microarrays (Moss *et al.*, 2018) or less often whole genome BS-seq (Allen Chan *et al.*, 2013). Although the results of

these studies were very promising, most of proposed biomarkers did not achieved early stage sensitivity and specificity required for real-word clinical applications.

It is crucial to anticipate that cfDNA is a complex mixture of DNA fragments released from various cells in the body and methylation changes observed in disease compared to healthy can be attributed to two mechanisms (Kustanovich *et al.*, 2019). First, aberrant methylation in a cancer tissue or cancer-induced epigenetic changes in other cell types can contribute to the altered methylation pattern in cfDNA. Second, methylation pattern can change due to different cell type contribution to cfDNA pool in disease. For instance, change in distribution of immune cell types could contribute to the altered methylation signal in circulating DNA. In many cases, it may be difficult to determine the exact source of aberrant methylation signal observed in cfDNA. For instance, genome-wide hypomethylation is one of the cancer hallmarks and it was captured in cfDNA of HCC patients (Allen Chan *et al.*, 2013). On the other hand, alterations in blood cells methylation patterns including progressing hypomethylation were also observed in liver cancer and are possible result of interplay between tumour and immune system (Koestler *et al.*, 2012; Y. Zhang *et al.*, 2018).

Although discrimination between tumour-specific methylation changes and changes derived from other cell-types in cancer may be challenging, all aberrant methylation signals could be used for cancer detection, regardless of their source, especially at the very early stages of cancerogenesis when ctDNA level is low. Moreover, study by Chen *et al.* suggests that restriction of methylation panel only to the tumour-specific signals does not improve cancer prediction (Chen *et al.*, 2020). In this study, they developed the PanSeer targeted methylation assay that included almost 477 regions identified based on tumour methylation data. However, around half of included loci exhibited different methylation status in tumour and plasma. These discrepancies could have been caused by

heterogeneity of methylation in an original tumour or by methylation signal dominated by cells types other than tumour. Surprisingly, discordant signals were important for cancer prediction since assay performance decreased from AUC = 0.99 to AUC = 0.97 after excluding these loci. This together suggest that liquid biopsy biomarkers identified directly from cfDNA rather than based on tumour-informed strategy could have better ability to detect cancer and overcome sensitivity issues in patients with low ctDNA level. Identification of methylation biomarkers directly from cfDNA requires comprehensive, whole-genome sequencing. However, this task is not trivial due to limited amount of cfDNA material in plasma. Whole genome bisulfite sequencing of cfDNA to high depth is technically feasible but requires high input DNA amount, has relatively low level of uniquely mapped, reduplicated reads and is not cost-effective (Allen Chan *et al.*, 2013; Worm Ørntoft *et al.*, 2017; M. C. Liu *et al.*, 2020). One of the most important objectives of this DPhil project was to address this technological gap and facilitate epigenetic studies of circulating DNA.

In the course of this project we developed two bisulfite-free methods: cfMeSeal and TAPS that can be easily applied to cfDNA.

CfMeSeal method is based on chemical and enzymatic labelling of 5mC and pull-down and enables enrichment-based sequencing of methylated regions. It can be considered as an alternative for antibody-based enrichment approaches. We demonstrated that this method is able to capture cancer-specific methylation changes in PDAC even at the low sequencing depth. Recent cfMeDIP-based studies confirmed that despite limited resolution, enrichment-based sequencing can provide methylation information sufficient for prediction of cancer and healthy (Shen *et al.*, 2018; Nassiri *et al.*, 2020). However, since the antibody-based enrichment method cfMeDIP has been developed based

on a protocol already used in the field, our method may not be as advantageous, especially as it requires TET enzyme, which has limited commercial availability.

In contrast, our second method, TAPS, is based on the novel chemistry and offers a unique approach for base-resolution methylation sequencing. TAPS has been already combined with long read sequencing to analyse epigenetic changes in mammalian and viral genomes (Y. Liu *et al.*, 2020). We have also successfully applied TAPS to cell-free DNA samples and single-cell TAPS is under development. Finally, we proposed cost-effective version of the method by combining TAPS with endonuclease treatment. Moreover, TAPS enabled development of new ways of bisulfite-free mapping of all modified cytosine variants. Contrary to other methods available in a field, BS-seq and EM-seq that convert unmodified cytosine, TAPS converts and introduces base-change directly in 5mC and 5hmC sites (Shapiro, DiFate and Welcher, 1974; Vaisvila *et al.*, 2019).

We proved that the TAPS method outperforms BS-seq in many aspects. As the TAPS protocol involves direct conversion and base-change of modified cytosines, it does not suffer from complexity issues and technical sequencing problems caused by BS-seq, has improved mapping rate and thus it is more cost effective. Its mild character enables its application to low input samples including cfDNA. The fact that TAPS converts only modified cytosines that are around 5% of all cytosines, enables development of further computational approaches enabling retrieving mutational information from TAPS data. Non-damaging character allows for simultaneous fragmentation analysis. Therefore, TAPS can provide multidimensional, complex information about cfDNA in a single sequencing assay. It is particularly important as sequencing costs are the most prohibitive factor in most of the studies. Integration of multiple layers of information is likely to help to develop diagnostic assay with excellent sensitivity and specificity and with real clinical

value. Although methylation can help for more sensitive cancer detection compared to mutations, overcome clonal mutagenesis problem and provide tool for cancer localisation in pan-cancer assays, we believe that it does not deny usefulness of mutational analysis. In particular, many cancer mutations are targeted by new generation cancer treatments and thus detection of actionable mutations in cfDNA could help to govern treatment decision (Turner *et al.*, 2020). Therefore, genetic and epigenetic cfDNA analysis should not be perceived as mutually exclusive approaches but rather as complementary methods that together can help to develop better diagnostic assays in the future. In this study we have shown that besides DNA methylation TAPS data provide reliable information about fragmentation profiles and that this information also have predictive potential. Although results of our analysis suggest that methylation alone enabled more sensitive and specific cancer detection compared to fragmentation patterns, small sample size does not allow for conclusive comparison of predictive performance of both approaches. Importantly, methylation and fragmentation features can be combined together to develop complex metaclassifiers with a potential for better prediction accuracy compared to the models based on single cfDNA feature. Computational approaches combining methylation and fragmentation features are currently under development. Finally, since most of nucleotides in TAPS data remain unchanged after conversion, we hope to extend our analysis of TAPS data to mutations, CNVs or other genomic features for comprehensive, single-assay genetic and epigenetic analysis of cfDNA. Such a combined approach could help to increase assay sensitivity to detect low levels of ctDNA even in shallow sequencing data.

Sensitivity limit to detect low level of ctDNA in cancer patients is one of the most important obstacles when developing early detection liquid biopsy assays and this issue should be carefully considered for TAPS. In this study, we used 10 ng of cfDNA to

prepare sequencing libraries in combination with relatively shallow, 12X sequencing. This amount of cfDNA can be purified from 1-4 mL of plasma, which is easy to obtain from most of the patients in clinical settings. However, low cfDNA input may be associated with limited sensitivity, especially when detecting early stage cancers with low level of tumour contribution (Bettegowda *et al.*, 2014). Assuming that molecular weight of haploid genome is around 3.3 pg, 10 ng cfDNA contains only 3000 haploid genome equivalents (Devonshire *et al.*, 2014). Since at early cancer stages tumour fraction in plasma is usually far below 1% (Bettegowda *et al.*, 2014), 10 ng of cfDNA is expected to contain less than 30 tumour-derived genomes. Therefore, detection of tumour-specific methylation signals using low DNA input may be challenging due to sensitivity limit caused by low abundance of ctDNA in the sample. First, rare methylation variants present in the original tumours can be absent in sequenced ~ 3000 genome equivalents due to the random sampling of plasma. Second, methylation changes present at the low frequencies in the sequencing libraries may go undetected at low or moderate sequencing depth such as 12X used in this study. Similarly, low cfDNA input and sequencing depth may not be optimal to detect rare mutation or CNVs with high confidence. On the other hand, TAPS provides genome-wide methylation information covering multiple loci and, therefore, has a higher probability of detecting tumour-derived aberrant methylation signals in patients with low tumour burden compared to assays based on small panels of analysed biomarkers. Recently, it was demonstrated that increased sequencing breadth and genome-wide mutational analysis could be more sensitive to detect low levels of ctDNA in cancer patients compared to deep targeted sequencing (Zviran, Rafael C. Schulman, *et al.*, 2020). Moreover, although TAPS was designed to analyse methylation, it can provide additional information such as cfDNA fragmentation patterns. Integrative, multidimensional cfDNA analysis combining

multiple types of biomarkers such as methylation marks, fragmentation profiles and CNVs may be another direction to improve sensitivity of TAPS to detect early stage disease at low sequencing depth and from low cfDNA input. Finally, aberrant methylation signal in cancer does not need to be limited to the tumour but can be potentially derived from other cell types including immune cells (Koestler *et al.*, 2012; Y. Zhang *et al.*, 2018). Detection of such a signal could overcome sensitivity limits in patients with low ctDNA amount. However, further studies are required to accurately estimate detection limits of TAPS in relation to the sequencing depth and cfDNA input. Specifically, parallel sequencing and comparative analysis of matched plasma-tumour samples as well as investigation of different cfDNA input amounts would allow for a precise assessment of TAPS sensitivity limit to detect low tumour fraction.

Multidimensional characterisation of cfDNA in a single sequencing could facilitate wider use of next generation sequencing in routine patient care. Although, whole-genome sequencing approaches provide the most comprehensive methods for genome and epigenome analysis, their application in the clinic is often debatable. Tumour sequencing has been proved to be useful in making treatment decision but it is not clear whether similar approaches can be used for early detection (Laskin *et al.*, 2015). The most common argument against whole-genome sequencing is cost. Schwarze *et al.* estimated sequencing costs of single cancer genome at 75X depth as £3420 (Schwarze *et al.*, 2019). Using this rate, 30X sequencing costs in clinical setting would be around £1368 or lower for assays that do not require high coverage. These costs may be prohibitive for a wide use of whole-genome sequencing in routine diagnostics. Moreover, data storage, data complexity and requirements for often sophisticated data analysis as well as potential difficulties in data interpretation may be considered as limitations that need to be considered before using whole-genome sequencing in standard medical practice. By

many, whole genome sequencing is still considered as a first step allowing for identification of useful biomarkers that will be combined into the final, simplified and cheaper assays. It is true that these assays will be cleaned from noisy signals and overall interpretation of data will be simplified but panel-based approach may also compromise overall assay performance as well. In contrary, some people envision that since sequencing costs are decreasing and expected to fall to \$1000 per genome in the near future, whole genome sequencing assays may become widely used in the medical practice (Wetterstrand KA, 2016). Novel, computational approaches can be standardised and automatized pipelines developed to facilitate integration and interpretation of generated data.

Another important fact is translation of scientific findings into clinical practice. Thus, it is important to demonstrate that biomarkers are reproducible between patients and that they offer significant benefit for a patient. Even when an assay shows high sensitivity and specificity compared to the traditional methods, its impact on early detection has to be considered. Early detection without improvement of survival or less invasive treatment may not be considered beneficial. Although desired clinical sensitivity and specificity should be close to 100% to avoid economic and social costs of misdiagnosis, the lowest values should be acceptable in some cases (Maxim, Niebo and Utell, 2014). The novel assays and approaches should be benchmarked to currently existing methods, their costs and overall invasiveness. For instance, even test with lower performance may be clinically useful when traditional diagnostic methods are worse or more expensive. Alternatively, lower performance can be also considered if a gold standard assay is very invasive. For instance, despite its worse performance compared to colonoscopy, Epi proColon has been approved for colorectal cancer screening owing to its lower invasiveness and better acceptance among patients (Shirley, 2020).

Unfortunately, accurate assessment of assay impact on patient survival is not trivial and ideally requires large-cohort studies with long-term monitoring of participants. On the other hand, most of studies, including ours, is performed on a post-diagnosis cohort of people who had already been diagnosed with cancer based on symptoms or by other methods. Even though results of our pilot study suggest a promising performance of cfDNA methylation analysis for cancer detection, our cohort did not represent early stage disease. Although most PDAC patients recruited for the study had resectable tumours without distant metastasis, almost all of them had regional disease with lymph node involvement. Tumour resection in pancreatic patients is associated with improved survival, however, overall prognosis are still poor with only 22% of patients surviving 5 years after diagnosis (Stark *et al.*, 2016; Neoptolemos *et al.*, 2018; Rawla, Sunkara and Gaduputi, 2019). Similarly, over 30% HCC patients included in the study had large tumours and were at stage III and IV. To note, less than 16% of people diagnosed at these stages survive 3 years after diagnosis (Abdel-Rahman, 2018). Since patients in our cohort do not represent early, curable disease, cancer detection at these stages may have minimal or no impact on prognosis and survival. Future efforts should be focused on validation of our findings in the patient groups with localised disease, small tumours and without lymph node involvement. Moreover, longitudinal studies, which involve long-term monitoring of patients with high risk, pre-malignant conditions e.g. cirrhosis, viral hepatitis, pancreatitis or pancreatic cysts could lead to development of diagnostic assays able for cancer detection in non-symptomatic individuals before clinical manifestation of disease. For instance, recent study published by Chen *et al.* tested ability of cancer detection in asymptomatic population using cfDNA methylation (Chen *et al.*, 2020). They collected blood from over 120 000 of healthy individuals and monitored their medical records for 8 years. They proved that they were able to predict cancer

development even prior 4 years before clinical diagnosis with AUC = 0.99 based on methylation biomarkers. Interestingly, this assay had worse performance on the symptomatic, cancer-diagnosed cohort, which may suggest that methylation biomarkers at early cancer development are different than those at more advanced cancer stages. Although, exact impact of earlier diagnosis on patient prognosis was not assessed in this study, this assay has ability to detect asymptomatic cancer cases and, in this way, contribute to cancer diagnosis at more curable stage, allow for an effective medical intervention and potentially have more significant impact on patient prognosis and survival. On the other hand, it is appreciated that such studies are very costly, logistically difficult and usually not available before assay is assessed in a pilot, symptomatic cohort. In our proof-of-concept study we demonstrated HCC AUC = 0.99 in non-cancer cohort which suggest that it may outperform many of currently used methods and proposed assays (Trevisani *et al.*, 2001; Allen Chan *et al.*, 2013; Cai *et al.*, 2019; Kisiel *et al.*, 2019; Sun *et al.*, 2019; Kotoh *et al.*, 2020). On the other hand, its performance on cirrhotic cohort needs to be further validated on bigger cohort to ensure our classifier is indeed able to discriminate well between chronic liver conditions and malignancy. Our assay performed slightly worse on non-metastatic PDAC with AUC = 0.96. On the other hand, this AUC is comparable or even slightly better than CA19-9 performance. Moreover, our assay showed promising discrimination between cancer and pancreatitis that still remain challenging with traditional methods (Goonetilleke and Siriwardena, 2007; Haab *et al.*, 2015; Cao *et al.*, 2016; Ngamruengphong and Canto, 2016). Further studies will be designed to validate our findings and further improve cancer detection based on cfDNA TAPS sequencing.

To summarise, this DPhil project provided novel tools for comprehensive analysis of cfDNA. In particular TAPS can be a novel tool allowing for complex analysis of

simultaneous epigenetic and genetic analysis of cfDNA. By applying TAPS to cfDNA from cancer and control individuals, we have demonstrated that information learned from this approach may have high predictive potential with overall performance close to that reported by other studies or even to approaches currently used in the clinic. Results of this study will be validated in the future on a bigger cohort and may serve as a starting point for other developments and real-world diagnostic assays based on cfDNA analysis. In a broader perspective this project may have positive impact on early cancer detection and patient prognosis.

Ithaca gave you the marvellous journey.

Without her you wouldn't have set out.

C. P. Cavafy *Ithaca*

8 Conclusions

In the course of this DPhil project I addressed all initial goals of my work. In particular, we managed to develop novel bisulfite-free methods for methylome sequencing and proved their usefulness for methylation analysis in clinical cfDNA samples.

The most important results and conclusions of this work are summarised below:

- We developed enrichment-based 5mC sequencing method by combining enzymatic and chemical labelling. We demonstrated that this protocol enables specific enrichment and sequencing of methylated cfDNA starting from as low as 5 ng of cfDNA. We demonstrated that even at 15M reads this method detects cancer-specific methylation changes in cfDNA with ability to distinguish between PDAC and healthy subjects.
- We developed novel chemistry to induce 5mC and 5hmC base change and apply it to develop base-resolution, bisulfite-free methylation sequencing method, TAPS. We confirmed that our method accurately detects 5mC and 5hmC in the genome and it overcomes major limitations of BS-seq. In particular, TAPS is not damaging and directly converts 5mC and 5hmC and it generates libraries with sequence complexity close to WGS. This is reflected in high mapping rate and improved sequencing quality.
- We optimised TAPS method for whole-genome, base-resolution sequencing of cfDNA. We demonstrated that TAPS generates high complexity sequencing libraries from as low as 10 ng of cfDNA and beside methylation, retains

information about cfDNA fragmentation patterns. We showed that both methylation and fragmentation patterns can be used for accurate detection of PDAC and HCC.

- We demonstrated that TAPS method can be modified to developed enrichment-based approaches. We enriched TAPS conversion product, DHU, by restriction digestion to develop eeTAPS. We confirmed that eeTAPS provides reliable methylation information consistent with wgTAPS and enables genome wide methylation sequencing at lower costs.

9 References

- Abdel-Rahman, O. (2018) ‘Assessment of the discriminating value of the 8th AJCC stage grouping for hepatocellular carcinoma’, *Hpb*. International Hepato-Pancreato-Biliary Association Inc., 20(1), pp. 41–48. doi: 10.1016/j.hpb.2017.08.017.
- Abdel-Wahab, O., Mullally, A., Hedvat, C., Garcia-Manero, G., Patel, J., Wadleigh, M., Malinge, S., Yao, J. J., Kilpivaara, O., Bhat, R., Huberman, K., Thomas, S., Dolgalev, I., Heguy, A., Paietta, E., Le Beau, M. M., Beran, M., Tallman, M. S., Ebert, B. L., Kantarjian, H. M., Stone, R. M., Gilliland, D. G., Crispino, J. D. and Levine, R. L. (2009) ‘Genetic characterization of TET1, TET2, and TET3 alterations in myeloid malignancies’, *Blood*, 114(1), pp. 144–147. doi: 10.1182/blood-2009-03-210039.
- Aberg, K. A., McClay, J. L., Nerella, S., Xie, L. Y., Clark, S. L., Hudson, A. D., Bukszár, J., Adkins, D., Consortium, S. S., Hultman, C. M., Sullivan, P. F., Magnusson, P. K. and van den Oord, E. J. C. G. J. (2012) ‘MBD-seq as a cost-effective approach for methylome-wide association studies: demonstration in 1500 case–control samples’, *Epigenomics*. Future Medicine Ltd London, UK, 4(6), pp. 605–621. doi: 10.2217/epi.12.59.
- Akalin, A., Garrett-Bakelman, F. E., Kormaksson, M., Busuttil, J., Zhang, L., Khrebtukova, I., Milne, T. A., Huang, Y., Biswas, D., Hess, J. L., Allis, C. D., Roeder, R. G., Valk, P. J. M., Löwenberg, B., Delwel, R., Fernandez, H. F., Paietta, E., Tallman, M. S., Schroth, G. P., Mason, C. E., Melnick, A. and Figueroa, M. E. (2012) ‘Base-pair resolution DNA methylation sequencing reveals profoundly divergent epigenetic landscapes in acute myeloid leukemia’, *PLoS Genetics*. Edited by H. L. Grimes. Public Library of Science, 8(6), p. e1002781. doi: 10.1371/journal.pgen.1002781.

Akkers, R. C., van Heeringen, S. J., Jacobi, U. G., Janssen-Megens, E. M., François, K. J., Stunnenberg, H. G. and Veenstra, G. J. C. (2009) 'A Hierarchy of H3K4me3 and H3K27me3 Acquisition in Spatial Gene Regulation in *Xenopus* Embryos', *Developmental Cell*, 17(3), pp. 425–434. doi: 10.1016/j.devcel.2009.08.005.

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A. J. R., Behjati, S., Biankin, A. V., Bignell, G. R., Bolli, N., Borg, A., Børresen-Dale, A. L., Boyault, S., Burkhardt, B., Butler, A. P., Caldas, C., Davies, H. R., Desmedt, C., Eils, R., Eyfjörd, J. E., Foekens, J. A., Greaves, M., Hosoda, F., Hutter, B., Illicic, T., Imbeaud, S., Imielinski, M., Jäger, N., Jones, D. T. W., Jonas, D., Knappskog, S., Koo, M., Lakhani, S. R., López-Otín, C., Martin, S., Munshi, N. C., Nakamura, H., Northcott, P. A., Pajic, M., Papaemmanuil, E., Paradiso, A., Pearson, J. V., Puente, X. S., Raine, K., Ramakrishna, M., Richardson, A. L., Richter, J., Rosenstiel, P., Schlesner, M., Schumacher, T. N., Span, P. N., Teague, J. W., Totoki, Y., Tutt, A. N. J., Valdés-Mas, R., Van Buuren, M. M., Van 't Veer, L., Vincent-Salomon, A., Waddell, N., Yates, L. R., Zucman-Rossi, J., Andrew Futreal, P., McDermott, U., Lichter, P., Meyerson, M., Grimmond, S. M., Siebert, R., Campo, E., Shibata, T., Pfister, S. M., Campbell, P. J. and Stratton, M. R. (2013) 'Signatures of mutational processes in human cancer', *Nature*. Nature Publishing Group, 500(7463), pp. 415–421. doi: 10.1038/nature12477.

Allen Chan, K. C., Jiang, P., Chan, C. W. M., Sun, K., Wong, J., Hui, E. P., Chan, S. L., Chan, W. C., Hui, D. S. C., Ng, S. S. M., Chan, H. L. Y., Wong, C. S. C., Ma, B. B. Y., Chan, A. T. C., Lai, P. B. S., Sun, H., Chiu, R. W. K. and Dennis Lo, Y. M. (2013) 'Noninvasive detection of cancer-associated genomewide hypomethylation and copy number aberrations by plasma DNA bisulfite sequencing', *Proceedings of the National Academy of Sciences of the United States of America*, 110(47), pp. 18761–18768. doi:

10.1073/pnas.1313995110.

Alliel, P. M., Seddiqi, N., Goudou, D., Cifuentes-Diaz, C., Romero, N., Velasco, E., Rieger, F. and Périn, J. P. (2000) 'Myoneurin, a novel member of the BTB/POZ-zinc finger family highly expressed in human muscle', *Biochemical and Biophysical Research Communications*. Academic Press Inc., 273(1), pp. 385–391. doi: 10.1006/bbrc.2000.2862.

Ally, A., Balasundaram, M., Carlsen, R., Chuah, E., Clarke, A., Dhalla, N., Holt, R. A., Jones, S. J. M., Lee, D., Ma, Y., Marra, M. A., Mayo, M., Moore, R. A., Mungall, A. J., Schein, J. E., Sipahimalani, P., Tam, A., Thiessen, N., Cheung, D., Wong, T., Brooks, D., Robertson, A. G., Bowlby, R., Mungall, K., Sadeghi, S., Xi, L., Covington, K., Shinbrot, E., Wheeler, D. A., Gibbs, R. A., Donehower, L. A., Wang, L., Bowen, J., Gastier-Foster, J. M., Gerken, M., Helsel, C., Leraas, K. M., Lichtenberg, T. M., Ramirez, N. C., Wise, L., Zmuda, E., Gabriel, S. B., Meyerson, M., Cibulskis, C., Murray, B. A., Shih, J., Beroukhi, R., Cherniack, A. D., Schumacher, S. E., Saksena, G., Pedamallu, C. S., Chin, L., Getz, G., Noble, M., Zhang, Hailei, Heiman, D., Cho, J., Gehlenborg, N., Saksena, G., Voet, D., Lin, P., Frazer, S., Defreitas, T., Meier, S., Lawrence, M., Kim, J., Creighton, C. J., Muzny, D., Doddapaneni, H. V., Hu, J., Wang, M., Morton, D., Korchina, V., Han, Y., Dinh, H., Lewis, L., Bellair, M., Liu, X., Santibanez, J., Glenn, R., Lee, S., Hale, W., Parker, J. S., Wilkerson, M. D., Hayes, D. N., Reynolds, S. M., Shmulevich, I., Zhang, W., Liu, Y., Iype, L., Makhlouf, H., Torbenson, M. S., Kakar, S., Yeh, M. M., Jain, D., Kleiner, D. E., Jain, D., Dhanasekaran, R., El-Serag, H. B., Yim, S. Y., Weinstein, J. N., Mishra, L., Zhang, Jianping, Akbani, R., Ling, S., Ju, Z., Su, X., Hegde, A. M., Mills, G. B., Lu, Y., Chen, J., Lee, J. S., Sohn, B. H., Shim, J. J., Tong, P., Aburatani, H., Yamamoto, S., Tatsuno, K., Li, W., Xia, Z., Stransky, N., Seiser, E.,

Innocenti, F., Gao, J., Kundra, R., Zhang, Hongxin, Heins, Z., Ochoa, A., Sander, C., Ladanyi, M., Shen, R., Arora, A., Sanchez-Vega, F., Schultz, N., Kasaian, K., Radenbaugh, A., Bissig, K. D., Moore, D. D., Totoki, Y., Nakamura, H., Shibata, T., Yau, C., Graim, K., Stuart, J., Haussler, D., Slagle, B. L., Ojesina, A. I., Katsonis, P., Koiré, A., Lichtarge, O., Hsu, T. K., Ferguson, M. L., Demchok, J. A., Felau, I., Sheth, M., Tarnuzzer, R., Wang, Z., Yang, L., Zenklusen, J. C., Zhang, Jiashan, Hutter, C. M., Sofia, H. J., Verhaak, R. G. W., Zheng, S., Lang, F., Chudamani, S., Liu, J., Lolla, L., Wu, Y., Naresh, R., Pihl, T., Sun, C., Wan, Y., Benz, C., Perou, A. H., Thorne, L. B., Boice, L., Huang, M., Rathmell, W. K., Noushmehr, H., Saggiaro, F. P., Tirapelli, D. P. da C., Junior, C. G. C., Mente, E. D., Silva, O. de C., Trevisan, F. A., Kang, K. J., Ahn, K. S., Giam, N. H., Moser, C. D., Giordano, T. J., Vinco, M., Welling, T. H., Crain, D., Curley, E., Gardner, J., Mallery, D., Morris, S., Paulauskis, J., Penny, R., Shelton, C., Shelton, T., Kelley, R., Park, J. W., Chandan, V. S., Roberts, L. R., Bathe, O. F., Hagedorn, C. H., Auman, J. T., O'Brien, D. R., Kocher, J. P. A., Jones, C. D., Mieczkowski, P. A., Perou, C. M., Skelly, T., Tan, D., Veluvolu, U., Balu, S., Bodenheimer, T., Hoyle, A. P., Jefferys, S. R., Meng, S., Mose, L. E., Shi, Y., Simons, J. V., Soloway, M. G., Roach, J., Hoadley, K. A., Baylin, S. B., Shen, H., Hinoue, T., Bootwalla, M. S., Van Den Berg, D. J., Weisenberger, D. J., Lai, P. H., Holbrook, A., Berrios, M. and Laird, P. W. (2017) 'Comprehensive and Integrative Genomic Characterization of Hepatocellular Carcinoma', *Cell*, 169(7), pp. 1327-1341.e23. doi: 10.1016/j.cell.2017.05.046.

Amemiya, H. M., Kundaje, A. and Boyle, A. P. (2019) 'The ENCODE Blacklist: Identification of Problematic Regions of the Genome', *Scientific Reports*. Nature Publishing Group, 9(1), pp. 1–5. doi: 10.1038/s41598-019-45839-z.

Argelaguet, R., Clark, S. J., Mohammed, H., Stapel, L. C., Krueger, C., Kapourani, C.

A., Imaz-Rosshandler, I., Lohoff, T., Xiang, Y., Hanna, C. W., Smallwood, S., Ibarra-Soria, X., Buettner, F., Sanguinetti, G., Xie, W., Krueger, F., Göttgens, B., Rugg-Gunn, P. J., Kelsey, G., Dean, W., Nichols, J., Stegle, O., Marioni, J. C. and Reik, W. (2019) ‘Multi-omics profiling of mouse gastrulation at single-cell resolution’, *Nature*. Nature Research, 576(7787), pp. 487–491. doi: 10.1038/s41586-019-1825-8.

Armstrong, K. M., Bermingham, E. N., Bassett, S. A., Treloar, B. P., Roy, N. C. and Barnett, M. P. (2011) ‘Global DNA methylation measurement by HPLC using low amounts of DNA’, *Biotechnology Journal*, 6(1), pp. 113–117. doi: 10.1002/biot.201000267.

Bailey, P., Chang, D. K., Nones, K., Johns, A. L., Patch, A. M., Gingras, M. C., Miller, D. K., Christ, A. N., Bruxner, T. J. C., Quinn, M. C., Nourse, C., Murtaugh, L. C., Harliwong, I., Idrisoglu, S., Manning, S., Nourbakhsh, E., Wani, S., Fink, L., Holmes, O., Chin, V., Anderson, M. J., Kazakoff, S., Leonard, C., Newell, F., Waddell, N. N., Wood, S., Xu, Q., Wilson, P. J., Cloonan, N., Kassahn, K. S., Taylor, D., Quek, K., Robertson, A., Pantano, L., Mincarelli, L., Sanchez, L. N., Evers, L., Wu, J., Pinese, M., Cowley, M. J., Jones, M. D., Colvin, E. K., Nagrial, A. M., Humphrey, E. S., Chantrill, L. A., Mawson, A., Humphris, J., Chou, A., Pajic, M., Scarlett, C. J., Pinho, A. V., Giry-Laterriere, M., Rooman, I., Samra, J. S., Kench, J. G., Lovell, J. A., Merrett, N. D., Toon, C. W., Epari, K., Nguyen, N. Q., Barbour, A., Zeps, N., Moran-Jones, K., Jamieson, N. B., Graham, J. S., Duthie, F., Oien, K., Hair, J., Grützmann, R., Maitra, A., Iacobuzio-Donahue, C. A., Wolfgang, C. L., Morgan, R. A., Lawlor, R. T., Corbo, V., Bassi, C., Rusev, B., Capelli, P., Salvia, R., Tortora, G., Mukhopadhyay, D., Petersen, G. M., Munzy, D. M., Fisher, W. E., Karim, S. A., Eshleman, J. R., Hruban, R. H., Pilarsky, C., Morton, J. P., Sansom, O. J., Scarpa, A., Musgrove, E. A., Bailey, U. M. H., Hofmann,

O., Sutherland, R. L., Wheeler, D. A., Gill, A. J., Gibbs, R. A., Pearson, J. V., Waddell, N. N., Biankin, A. V. and Grimmond, S. M. (2016) 'Genomic analyses identify molecular subtypes of pancreatic cancer', *Nature*. Nature Publishing Group, 531(7592), pp. 47–52. doi: 10.1038/nature16965.

Bannister, A. J. and Kouzarides, T. (2011) 'Regulation of chromatin by histone modifications', *Cell Research*. Nature Publishing Group, pp. 381–395. doi: 10.1038/cr.2011.22.

Bararia, A., Dey, S., Gulati, S., Ghatak, S., Ghosh, S., Banerjee, S. and Sikdar, N. (2020) 'Differential methylation landscape of pancreatic ductal adenocarcinoma and its precancerous lesions', *Hepatobiliary and Pancreatic Diseases International*. Elsevier B.V., 19(3), pp. 205–217. doi: 10.1016/j.hbpd.2020.03.010.

Beard, C., Li, E. and Jaenisch, R. (1995) 'Loss of methylation activates Xist in somatic but not in embryonic cells', *Genes and Development*. Cold Spring Harbor Laboratory Press, 9(19), pp. 2325–2334. doi: 10.1101/gad.9.19.2325.

Benzing, C., Lam, H., Tsang, C. M., Rimmer, A., Arroyo-Berdugo, Y., Calle, Y. and Wells, C. M. (2019) 'TIMP-2 secreted by monocyte-like cells is a potent suppressor of invadopodia formation in pancreatic cancer cells', *BMC Cancer*. BioMed Central Ltd., 19(1), p. 1214. doi: 10.1186/s12885-019-6429-z.

Berdasco, M. and Esteller, M. (2018) 'Clinical epigenetics: seizing opportunities for translation', *Nature Reviews Genetics*. Nature Publishing Group, 20(2), pp. 109–127. doi: 10.1038/s41576-018-0074-2.

Bernstein, B. E., Kamal, M., Lindblad-Toh, K., Bekiranov, S., Bailey, D. K., Huebert, D. J., McMahon, S., Karlsson, E. K., Kulbokas, E. J., Gingeras, T. R., Schreiber, S. L. and

Lander, E. S. (2005) ‘Genomic maps and comparative analysis of histone modifications in human and mouse’, *Cell*. Elsevier, 120(2), pp. 169–181. doi: 10.1016/j.cell.2005.01.001.

Bernstein, B. E., Meissner, A. and Lander, E. S. (2007) ‘The Mammalian Epigenome’, *Cell*. Elsevier, pp. 669–681. doi: 10.1016/j.cell.2007.01.033.

Bernstein, B. E., Mikkelsen, T. S., Xie, X., Kamal, M., Huebert, D. J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K., Jaenisch, R., Wagschal, A., Feil, R., Schreiber, S. L. and Lander, E. S. (2006) ‘A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells’, *Cell*. Elsevier, 125(2), pp. 315–326. doi: 10.1016/j.cell.2006.02.041.

Best, M. G., Sol, N., Tannous, B. A., Wesseling, P. and Wurdinger, T. (2015) ‘RNA-Seq of Tumor-Educated Platelets Enables Blood-Based Pan-Cancer, Multiclass, and Molecular Pathway Cancer Diagnostics’, *Cancer Cell*, 28, pp. 666–676. doi: 10.1016/j.ccell.2015.09.018.

Bettegowda, C., Sausen, M., Leary, R. J., Kinde, I., Wang, Y., Agrawal, N., Bartlett, B. R., Wang, H., Lubner, B., Alani, R. M., Antonarakis, E. S., Azad, N. S., Bardelli, A., Brem, H., Cameron, J. L., Lee, C. C., Fecher, L. A., Gallia, G. L., Gibbs, P., Le, D., Giuntoli, R. L., Goggins, M., Hogarty, M. D., Holdhoff, M., Hong, S. M. S.-M., Jiao, Y., Juhl, H. H., Kim, J. J., Siravegna, G., Laheru, D. A., Lauricella, C., Lim, M., Lipson, E. J., Marie, S. K. N., Netto, G. J., Oliner, K. S., Olivi, A., Olsson, L., Riggins, G. J., Sartore-Bianchi, A., Schmidt, K., Shih, I. M., Oba-Shinjo, S. M., Siena, S., Theodorescu, D., Tie, J., Harkins, T. T., Veronese, S., Wang, T. L. T.-L., Weingart, J. D., Wolfgang, C. L., Wood, L. D., Xing, D., Hruban, R. H., Wu, J., Allen, P. J., Schmidt, C. M., Choti, M. A., Velculescu, V. E., Kinzler, K. W., Vogelstein, B., Papadopoulos, N., Diaz, L. A.,

Agrawal, N., Kim, J. J., Lauricella, C., Xing, D., Papadopoulos, N., Schmidt, C. M., Wu, J., Jiao, Y., Wolfgang, C. L., Shih, L.-M., Oba-Shinjo, S. M., Hong, S. M. S.-M., Schmidt, K., Harkins, T. T., Kinde, I., Bartlett, B. R., Siena, S., Netto, G. J., Choti, M. A., Theodorescu, D., Fecher, L. A., Lubner, B., Riggins, G. J., Lee, C. C., Alani, R. M., Wang, T. L. T.-L., Oliner, K. S., Azad, N. S., Wang, H., Cameron, J. L., Bardelli, A., Gallia, G. L., Gibbs, P., Velculescu, V. E., Bettegowda, C., Sausen, M., Giuntoli, R. L., Vogelstein, B., Goggins, M., Holdhoff, M., Brem, H., Le, D., Lipson, E. J., Tie, J., Leary, R. J., Lim, M., Weingart, J. D., Veronese, S., Siravegna, G., Juhl, H. H., Kinzler, K. W., Olsson, L., Sartore-Bianchi, A., Hruban, R. H., Wang, Y., Marie, S. K. N., Laheru, D. A., Diaz, L. A., Antonarakis, E. S., Wood, L. D., Hogarty, M. D. and Allen, P. J. (2014) 'Detection of Circulating Tumor DNA in Early- and Late-Stage Human Malignancies', *Science Translational Medicine*. NIH Public Access, 6(224), pp. 224ra24-224ra24. doi: 10.1126/scitranslmed.3007094.

Bhattacharyya, S., Pradhan, K., Campbell, N., Mazdo, J., Vasantkumar, A., Maqbool, S., Bhagat, T. D., Gupta, S., Suzuki, M., Yu, Y., Grealley, J. M., Steidl, U., Bradner, J., Dawlaty, M., Godley, L., Maitra, A. and Verma, A. (2017) 'Altered hydroxymethylation is seen at regulatory regions in pancreatic cancer and regulates oncogenic pathways', *Genome Research*. Cold Spring Harbor Laboratory Press, 27(11), pp. 1830–1842. doi: 10.1101/gr.222794.117.

Bogu, G. K., Vizán, P., Stanton, L. W., Beato, M., Di Croce, L. and Marti-Renom, M. A. (2016) 'Chromatin and RNA Maps Reveal Regulatory Long Noncoding RNAs in Mouse', *Molecular and Cellular Biology*. American Society for Microbiology, 36(5), pp. 809–819. doi: 10.1128/mcb.00955-15.

Bonev, B. and Cavalli, G. (2016) 'Organization and function of the 3D genome', *Nature*

Reviews Genetics. Nature Publishing Group, pp. 661–678. doi: 10.1038/nrg.2016.112.

Booth, M. J., Branco, M. R., Ficz, G., Oxley, D., Krueger, F., Reik, W. and Balasubramanian, S. (2012) ‘Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution.’, *Science (New York, N.Y.)*. American Association for the Advancement of Science, 336(6083), pp. 934–7. doi: 10.1126/science.1220671.

Booth, M. J., Marsico, G., Bachman, M., Beraldi, D. and Balasubramanian, S. (2014) ‘Quantitative sequencing of 5-formylcytosine in DNA at single-base resolution’, *Nature Chemistry*. Nature Publishing Group, 6(5), pp. 435–440. doi: 10.1038/nchem.1893.

Booth, M. J., Ost, T. W. B., Beraldi, D., Bell, N. M., Branco, M. R., Reik, W. and Balasubramanian, S. (2013) ‘Oxidative bisulfite sequencing of 5-methylcytosine and 5-hydroxymethylcytosine’, *Nature Protocols*, 8(10), pp. 1841–1851. doi: 10.1038/nprot.2013.115.

Borgel, J., Guibert, S., Li, Y., Chiba, H., Schübeler, D., Sasaki, H., Forné, T. and Weber, M. (2010) ‘Targets and dynamics of promoter DNA methylation during early mouse development’, *Nature Genetics*. Nat Genet, 42(12), pp. 1093–1100. doi: 10.1038/ng.708.

Bourc’his, D., Xu, G. L., Lin, C. S., Bollman, B. and Bestor, T. H. (2001) ‘Dnmt3L and the establishment of maternal genomic imprints’, *Science*, 294(5551), pp. 2536–2539. doi: 10.1126/science.1065848.

Brancaccio, M., Natale, F., Falco, G. and Angrisano, T. (2020) ‘Cell-free dna methylation: The new frontiers of pancreatic cancer biomarkers’ discovery’, *Genes*, p. 14. doi: 10.3390/genes11010014.

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A. and Jemal, A. (2018) ‘Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries’, *CA: A Cancer Journal for Clinicians*. Wiley, 68(6), pp. 394–424. doi: 10.3322/caac.21492.

Breiling, A. and Lyko, F. (2015) *Epigenetic regulatory functions of DNA modifications: 5-methylcytosine and beyond*, *Epigenetics and Chromatin*. doi: 10.1186/s13072-015-0016-6.

Bucknall, M., Fung, K. Y. C. and Duncan, M. W. (2002) ‘Practical quantitative biomedical applications of MALDI-TOF mass spectrometry’, *Journal of the American Society for Mass Spectrometry*. No longer published by Elsevier, 13(9), pp. 1015–1027. doi: 10.1016/S1044-0305(02)00426-9.

Cadieux, B., Ching, T. T., VandenBerg, S. R. and Costello, J. F. (2006) ‘Genome-wide hypomethylation in human glioblastomas associated with specific copy number alteration, methylenetetrahydrofolate reductase allele status, and increased proliferation’, *Cancer Research*, 66(17), pp. 8469–8476. doi: 10.1158/0008-5472.CAN-06-1547.

Cai, J., Chen, L., Zhang, Z., Zhang, Xinyu, Lu, X., Liu, W., Shi, G., Ge, Y., Gao, P., Yang, Y., Ke, A., Xiao, L., Dong, R., Zhu, Y., Yang, X., Wang, J., Zhu, T., Yang, D., Huang, X., Sui, C., Qiu, S., Shen, F., Sun, H., Zhou, W., Zhou, J., Nie, J., Zeng, C., Stroup, E. K., Zhang, Xu, Chiu, B. C.-H., Lau, W. Y., He, C., Wang, H., Zhang, W. and Fan, J. (2019) ‘Genome-wide mapping of 5-hydroxymethylcytosines in circulating cell-free DNA as a non-invasive approach for early detection of hepatocellular carcinoma’, *Gut*. BMJ Publishing Group, 68(12), pp. 2195–2205. doi: 10.1136/gutjnl-2019-318882.

Cancer Research UK (2020) *Worldwide cancer statistics* | *Cancer Research UK*, *Cancer*

Research UK. Available at: <https://www.cancerresearchuk.org/health-professional/cancer-statistics/worldwide-cancer> (Accessed: 11 June 2020).

Cao, S., Hu, Y., Gao, X., Liao, Q. and Zhao, Y. (2016) 'Serum carbohydrate antigen 19-9 in differential diagnosis of benign and malignant pancreatic cystic neoplasms: A meta-Analysis', *PLoS ONE*, 11(11), pp. 1–13. doi: 10.1371/journal.pone.0166406.

Chabchoub, G., Uz, E., Maalej, A., Mustafa, C. A., Rebai, A., Mnif, M., Bahloul, Z., Farid, N. R., Ozcelik, T. and Ayadi, H. (2009) 'Analysis of skewed X-chromosome inactivation in females with rheumatoid arthritis and autoimmune thyroid diseases', *Arthritis Research and Therapy*. BioMed Central, 11(4), p. R106. doi: 10.1186/ar2759.

Chabon, J. J., Simmons, A. D., Lovejoy, A. F., Esfahani, M. S., Newman, A. M., Haringsma, H. J., Kurtz, D. M., Stehr, H., Scherer, F., Karlovich, C. A., Harding, T. C., Durkin, K. A., Otterson, G. A., Purcell, W. T., Camidge, D. R., Goldman, J. W., Sequist, L. V., Piotrowska, Z., Wakelee, H. A., Neal, J. W., Alizadeh, A. A. and Diehn, M. (2016) 'Circulating tumour DNA profiling reveals heterogeneity of EGFR inhibitor resistance mechanisms in lung cancer patients', *Nature Communications*, 7. doi: 10.1038/ncomms11815.

Chen, E. Y., Tan, C. M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G. V., Clark, N. R. and Ma'ayan, A. (2013) 'Enrichr: Interactive and collaborative HTML5 gene list enrichment analysis tool', *BMC Bioinformatics*. BMC Bioinformatics, 14. doi: 10.1186/1471-2105-14-128.

Chen, T., Hevi, S., Gay, F., Tsujimoto, N., He, T., Zhang, B., Ueda, Y. and Li, E. (2007) 'Complete inactivation of DNMT1 leads to mitotic catastrophe in human cancer cells', *Nature Genetics*. Nature Publishing Group, 39(3), pp. 391–396. doi: 10.1038/ng1982.

Chen, X., Gole, J., Gore, A., He, Q., Lu, M., Min, J., Yuan, Z., Yang, X., Jiang, Y., Zhang, T., Suo, C., Li, X., Cheng, L., Zhang, Z., Niu, H., Li, Z., Xie, Z., Shi, H., Zhang, X., Fan, M., Wang, X., Yang, Y., Dang, J., McConnell, C., Zhang, J., Wang, J., Yu, S., Ye, W., Gao, Y., Zhang, K., Liu, R. and Jin, L. (2020) ‘Non-invasive early detection of cancer four years before conventional diagnosis using a blood test’, *Nature Communications*. Nature Publishing Group, 11(1), p. 3475. doi: 10.1038/s41467-020-17316-z.

Chowdhury, B., Cho, I. H., Hahn, N. and Irudayaraj, J. (2014) ‘Quantification of 5-methylcytosine, 5-hydroxymethylcytosine and 5-carboxylcytosine from the blood of cancer patients by an enzyme-based immunoassay’, *Analytica Chimica Acta*. Elsevier, 852, pp. 212–217. doi: 10.1016/j.aca.2014.09.020.

Cohen, J. D., Javed, A. A., Thoburn, C., Wong, F., Tie, J., Gibbs, P., Schmidt, C. M., Yip-Schneider, M. T., Allen, P. J., Schattner, M., Brand, R. E., Singhi, A. D., Petersen, G. M., Hong, S.-M., Kim, S. C., Falconi, M., Doglioni, C., Weiss, M. J., Ahuja, N., He, J., Makary, M. A., Maitra, A., Hanash, S. M., Dal Molin, M., Wang, Y., Li, L., Ptak, J., Dobbyn, L., Schaefer, J., Silliman, N., Popoli, M., Goggins, M. G., Hruban, R. H., Wolfgang, C. L., Klein, A. P., Tomasetti, C., Papadopoulos, N., Kinzler, K. W., Vogelstein, B. and Lennon, A. M. (2017) ‘Combined circulating tumor DNA and protein biomarker-based liquid biopsy for the earlier detection of pancreatic cancers’, *Proceedings of the National Academy of Sciences*. National Academy of Sciences, 114(38), pp. 10202–10207. doi: 10.1073/pnas.1704961114.

Cohen, J. D., Javed, A. A., Thoburn, C., Wong, F., Tie, J., Gibbs, P., Schmidt, C. M., Yip-Schneider, M. T., Allen, P. J., Schattner, M., Brand, R. E., Singhi, A. D., Petersen, G. M., Hong, S.-M., Kim, S. C., Falconi, M., Doglioni, C., Weiss, M. J., Ahuja, N., He, J., Makary, M. A., Maitra, A., Hanash, S. M., Dal Molin, M., Wang, Y., Li, L., Ptak, J.,

Dobbyn, L., Schaefer, J., Silliman, N., Popoli, M., Goggins, M. G., Hruban, R. H., Wolfgang, C. L., Klein, A. P., Tomasetti, C., Papadopoulos, N., Kinzler, K. W., Vogelstein, B., Lennon, A. M., Javed, A. A., Schaefer, J., Hong, S.-M., Ahuja, N., Lennon, A. M., Schattner, M., Singhi, A. D., Hruban, R. H., Li, L., Petersen, G. M., Kim, S. C., Wolfgang, C. L., Klein, A. P., Brand, R. E., Wong, F., Schmidt, C. M., Falconi, M., Papadopoulos, N., He, J., Maitra, A., Yip-Schneider, M. T., Tomasetti, C., Ptak, J., Cohen, J. D., Hanash, S. M., Vogelstein, B., Wang, Y., Silliman, N., Goggins, M. G., Tie, J., Weiss, M. J., Thoburn, C., Dobbyn, L., Gibbs, P., Doglioni, C., Kinzler, K. W., Makary, M. A., Allen, P. J., Dal Molin, M., Javed, A. A., Thoburn, C., Wong, F., Tie, J., Gibbs, P., Schmidt, C. M., Yip-Schneider, M. T., Allen, P. J., Schattner, M., Brand, R. E., Singhi, A. D., Petersen, G. M., Hong, S.-M., Kim, S. C., Falconi, M., Doglioni, C., Weiss, M. J., Ahuja, N., He, J., Makary, M. A., Maitra, A., Hanash, S. M., Dal Molin, M., Wang, Y., Li, L., Ptak, J., Dobbyn, L., Schaefer, J., Silliman, N., Popoli, M., Goggins, M. G., Hruban, R. H., Wolfgang, C. L., Klein, A. P., Tomasetti, C., Papadopoulos, N., Kinzler, K. W., Vogelstein, B. and Lennon, A. M. (2017) ‘Combined circulating tumor DNA and protein biomarker-based liquid biopsy for the earlier detection of pancreatic cancers’, *Proceedings of the National Academy of Sciences*. National Academy of Sciences, 114(38), pp. 10202–10207. doi: 10.1073/pnas.1704961114.

Cohen, J. D., Li, L., Wang, Y., Thoburn, C., Afsari, B., Danilova, L., Douville, C., Javed, A. A., Wong, F., Mattox, A., Hruban, R. H., Wolfgang, C. L., Goggins, M. G., Molin, M. D., Wang, T. L., Roden, R., Klein, A. P., Ptak, J., Dobbyn, L., Schaefer, J., Silliman, N., Popoli, M., Vogelstein, J. T., Browne, J. D., Schoen, R. E., Brand, R. E., Tie, J., Gibbs, P., Wong, H. L., Mansfield, A. S., Jen, J., Hanash, S. M., Falconi, M., Allen, P. J., Zhou, S., Bettegowda, C., Diaz, L. A., Tomasetti, C., Kinzler, K. W., Vogelstein, B., Lennon, A. M. and Papadopoulos, N. (2018) ‘Detection and localization of surgically resectable

cancers with a multi-analyte blood test’, *Science*. American Association for the Advancement of Science, 359(6378), pp. 926–930. doi: 10.1126/science.aar3247.

Cotton, A. M., Price, E. M., Jones, M. J., Balaton, B. P., Kobor, M. S. and Brown, C. J. (2015) ‘Landscape of DNA methylation on the X chromosome reflects CpG density, functional chromatin state and X-chromosome inactivation’, *Human Molecular Genetics*. Oxford University Press, 24(6), pp. 1528–1539. doi: 10.1093/hmg/ddu564.

Creyghton, M. P., Cheng, A. W., Welstead, G. G., Kooistra, T., Carey, B. W., Steine, E. J., Hanna, J., Lodato, M. A., Frampton, G. M., Sharp, P. A., Boyer, L. A., Young, R. A. and Jaenisch, R. (2010) ‘Histone H3K27ac separates active from poised enhancers and predicts developmental state’, *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 107(50), pp. 21931–21936. doi: 10.1073/pnas.1016071107.

Cristiano, S., Leal, A., Phallen, J., Fiksel, J., Adleff, V., Bruhm, D. C., Jensen, S. Ø., Medina, J. E., Hruban, C., White, J. R., Palsgrove, D. N., Niknafs, N., Anagnostou, V., Forde, P., Naidoo, J., Marrone, K., Brahmer, J., Woodward, B. D., Husain, H., van Rooijen, K. L., Ørntoft, M. B. W., Madsen, A. H., van de Velde, C. J. H., Verheij, M., Cats, A., Punt, C. J. A., Vink, G. R., van Grieken, N. C. T., Koopman, M., Fijneman, R. J. A., Johansen, J. S., Nielsen, H. J., Meijer, G. A., Andersen, C. L., Scharpf, R. B. and Velculescu, V. E. (2019) *Genome-wide cell-free DNA fragmentation in patients with cancer*, *Nature*. Nature Publishing Group. doi: 10.1038/s41586-019-1272-6.

Csankovszki, G., Nagy, A. and Jaenisch, R. (2001) ‘Synergism of Xist RNA, DNA methylation, and histone hypoacetylation in maintaining X chromosome inactivation’, *Journal of Cell Biology*. The Rockefeller University Press, 153(4), pp. 773–783. doi: 10.1083/jcb.153.4.773.

Devonshire, A. S., Whale, A. S., Gutteridge, A., Jones, G., Cowen, S., Foy, C. A. and Huggett, J. F. (2014) 'Towards standardisation of cell-free DNA measurement in plasma: Controls for extraction efficiency, fragment size bias and quantification', *Analytical and Bioanalytical Chemistry*, 406(26), pp. 6499–6512. doi: 10.1007/s00216-014-7835-3.

Diehl, F., Li, M., Dressman, D., He, Y., Shen, D., Szabo, S., Diaz, L. A., Goodman, S. N., David, K. A., Juhl, H., Kinzler, K. W. and Vogelstein, B. (2005) 'Detection and quantification of mutations in the plasma of patients with colorectal tumors', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 102(45), pp. 16368–16373. doi: 10.1073/pnas.0507904102.

Diehl, F., Schmidt, K., Choti, M. A., Romans, K., Goodman, S., Li, M., Thornton, K., Agrawal, N., Sokoll, L., Szabo, S. A., Kinzler, K. W., Vogelstein, B. and Diaz, L. A. (2008) 'Circulating mutant DNA to assess tumor dynamics', *Nature Medicine*. Nature Publishing Group, 14(9), pp. 985–990. doi: 10.1038/nm.1789.

van Dijk, E. L., Jaszczyszyn, Y., Naquin, D. and Thermes, C. (2018) 'The Third Revolution in Sequencing Technology', *Trends in Genetics*. Elsevier Ltd, 34(9), pp. 666–681. doi: 10.1016/j.tig.2018.05.008.

Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. A., Doyle, F., Epstein, C. B., Frietze, S., Harrow, J., Kaul, R., Khatun, J., Lajoie, B. R., Landt, S. G., Lee, B. K., Pauli, F., Rosenbloom, K. R., Sabo, P., Safi, A., Sanyal, A., Shores, N., Simon, J. M., Song, L., Trinklein, N. D., Altshuler, R. C., Birney, E., Brown, J. B., Cheng, C., Djebali, S., Dong, X., Ernst, J., Furey, T. S., Gerstein, M., Giardine, B., Greven, M., Hardison, R. C., Harris, R. S., Herrero, J., Hoffman, M. M., Iyer, S., Kellis, M., Kheradpour, P., Lassmann, T., Li, Q., Lin, X., Marinov, G. K., Merkel, A., Mortazavi, A., Parker, S. C. J., Reddy, T. E., Rozowsky, J., Schlesinger, F., Thurman, R. E., Wang, J., Ward, L. D.,

Whitfield, T. W., Wilder, S. P., Wu, W., Xi, H. S., Yip, K. Y., Zhuang, J., Bernstein, B. E., Green, E. D., Gunter, C., Snyder, M., Pazin, M. J., Lowdon, R. F., Dillon, L. A. L., Adams, L. B., Kelly, C. J., Zhang, J., Wexler, J. R., Good, P. J., Feingold, E. A., Crawford, G. E., Dekker, J., Elnitski, L., Farnham, P. J., Giddings, M. C., Gingeras, T. R., Guigó, R., Hubbard, T. J., Kent, W. J., Lieb, J. D., Margulies, E. H., Myers, R. M., Stamatoyannopoulos, J. A., Tenenbaum, S. A., Weng, Z., White, K. P., Wold, B., Yu, Y., Wrobel, J., Risk, B. A., Gunawardena, H. P., Kuiper, H. C., Maier, C. W., Xie, L., Chen, X., Mikkelsen, T. S., Gillespie, S., Goren, A., Ram, O., Zhang, X., Wang, L., Issner, R., Coyne, M. J., Durham, T., Ku, M., Truong, T., Eaton, M. L., Dobin, A., Tanzer, A., Lagarde, J., Lin, W., Xue, C., Williams, B. A., Zaleski, C., Röder, M., Kokocinski, F., Abdelhamid, R. F., Alioto, T., Antoshechkin, I., Baer, M. T., Batut, P., Bell, I., Bell, K., Chakraborty, S., Chrast, J., Curado, J., Derrien, T., Drenkow, J., Dumais, E., Dumais, J., Duttagupta, R., Fastuca, M., Fejes-Toth, K., Ferreira, P., Foissac, S., Fullwood, M. J., Gao, H., Gonzalez, D., Gordon, A., Howald, C., Jha, S., Johnson, R., Kapranov, P., King, B., Kingswood, C., Li, G., Luo, O. J., Park, E., Preall, J. B., Presaud, K., Ribeca, P., Robyr, D., Ruan, X., Sammeth, M., Sandhu, K. S., Schaeffer, L., See, L. H., Shahab, A., Skancke, J., Suzuki, A. M., Takahashi, H., Tilgner, H., Trout, D., Walters, N., Wang, Huaian, Hayashizaki, Y., Reymond, A., Antonarakis, S. E., Hannon, G. J., Ruan, Y., Carninci, P., Sloan, C. A., Learned, K., Malladi, V. S., Wong, M. C., Barber, G. P., Cline, M. S., Dreszer, T. R., Heitner, S. G., Karolchik, D., Kirkup, V. M., Meyer, L. R., Long, J. C., Maddren, M., Raney, B. J., Grasfeder, L. L., Giresi, P. G., Battenhouse, A., Sheffield, N. C., Showers, K. A., London, D., Bhinge, A. A., Shestak, C., Schaner, M. R., Kim, S. K., Zhang, Z. Z., Mieczkowski, P. A., Mieczkowska, J. O., Liu, Z., McDaniell, R. M., Ni, Y., Rashid, N. U., Kim, M. J., Adar, S., Zhang, Zhancheng, Wang, T., Winter, D., Keefe, D., Iyer, V. R., Zheng, M., Wang, P., Gertz, J., Vielmetter, J.,

Partridge, E. C., Varley, K. E., Gasper, C., Bansal, A., Pepke, S., Jain, P., Amrhein, H., Bowling, K. M., Anaya, M., Cross, M. K., Muratet, M. A., Newberry, K. M., McCue, K., Nesmith, A. S., Fisher-Aylor, K. I., Pusey, B., DeSalvo, G., Parker, S. L., Balasubramanian, Sreeram, Davis, N. S., Meadows, S. K., Eggleston, T., Newberry, J. S., Levy, S. E., Absher, D. M., Wong, W. H., Blow, M. J., Visel, A., Pennachio, L. A., Petrykowska, H. M., Abyzov, A., Aken, B., Barrell, D., Barson, G., Berry, A., Bignell, A., Boychenko, V., Bussotti, G., Davidson, C., Despacio-Reyes, G., Diekhans, M., Ezkurdia, I., Frankish, A., Gilbert, J., Gonzalez, J. M., Griffiths, E., Harte, R., Hendrix, D. A., Hunt, T., Jungreis, I., Kay, M., Khurana, E., Leng, J., Lin, M. F., Loveland, J., Lu, Z., Manthravadi, D., Mariotti, M., Mudge, J., Mukherjee, G., Notredame, C., Pei, B., Rodriguez, J. M., Saunders, G., Sboner, A., Searle, S., Sisu, C., Snow, C., Steward, C., Tapanari, E., Tress, M. L., Van Baren, M. J., Washietl, S., Wilming, L., Zadissa, A., Zhang, Zhengdong, Brent, M., Haussler, D., Valencia, A., Addleman, N., Alexander, R. P., Auerbach, R. K., Balasubramanian, Suganthi, Bettinger, K., Bhardwaj, N., Boyle, A. P., Cao, A. R., Cayting, P., Charos, A., Cheng, Y., Eastman, C., Euskirchen, G., Fleming, J. D., Grubert, F., Habegger, L., Hariharan, M., Harmanci, A., Iyengar, S., Jin, V. X., Karczewski, K. J., Kasowski, M., Lacroute, P., Lam, H., Lamarre-Vincent, N., Lian, J., Lindahl-Allen, M., Min, R., Miotto, B., Monahan, H., Moqtaderi, Z., Mu, X. J., O'Geen, H., Ouyang, Z., Patacsil, D., Raha, D., Ramirez, L., Reed, B., Shi, M., Slifer, T., Witt, H., Wu, L., Xu, X., Yan, K. K., Yang, X., Struhl, K., Weissman, S. M., Penalva, L. O., Karmakar, S., Bhanvadia, R. R., Choudhury, A., Domanus, M., Ma, L., Moran, J., Victorsen, A., Auer, T., Centanin, L., Eichenlaub, M., Gruhl, F., Heermann, S., Hoeckendorf, B., Inoue, D., Kellner, T., Kirchmaier, S., Mueller, C., Reinhardt, R., Schertel, L., Schneider, S., Sinn, R., Wittbrodt, B., Wittbrodt, J., Jain, G., Balasundaram, G., Bates, D. L., Byron, R., Canfield, T. K., Diegel, M. J., Dunn, D., Ebersol, A. K.,

Frum, T., Garg, K., Gist, E., Hansen, R. S., Boatman, L., Haugen, E., Humbert, R., Johnson, A. K., Johnson, E. M., Kutyavin, T. V., Lee, K., Lotakis, D., Maurano, M. T., Neph, S. J., Neri, F. V., Nguyen, E. D., Qu, H., Reynolds, A. P., Roach, V., Rynes, E., Sanchez, M. E., Sandstrom, R. S., Shafer, A. O., Stergachis, A. B., Thomas, S., Vernot, B., Vierstra, J., Vong, S., Wang, Hao, Weaver, M. A., Yan, Y., Zhang, M., Akey, J. M., Bender, M., Dorschner, M. O., Groudine, M., MacCoss, M. J., Navas, P., Stamatoyannopoulos, G., Beal, K., Brazma, A., Flicek, P., Johnson, N., Lusk, M., Luscombe, N. M., Sobral, D., Vaquerizas, J. M., Batzoglou, S., Sidow, A., Hussami, N., Kyriazopoulou-Panagiotopoulou, S., Libbrecht, M. W., Schaub, M. A., Miller, W., Bickel, P. J., Banfai, B., Boley, N. P., Huang, H., Li, J. J., Noble, W. S., Bilmes, J. A., Buske, O. J., Sahu, A. D., Kharchenko, P. V., Park, P. J., Baker, D., Taylor, J. and Lochovsky, L. (2012) ‘An integrated encyclopedia of DNA elements in the human genome’, *Nature*. Nature Publishing Group, 489(7414), pp. 57–74. doi: 10.1038/nature11247.

Ehrlich, M. (2009) ‘DNA hypomethylation in cancer cells’, *Epigenomics*. NIH Public Access, pp. 239–259. doi: 10.2217/EPI.09.33.

Ehrlich, M., Jiang, G., Fiala, E., Dome, J. S., Yu, M. C., Long, T. I., Youn, B., Sohn, O.-S., Widschwendter, M., Tomlinson, G. E., Chintagumpala, M., Champagne, M., Parham, D., Liang, G., Malik, K. and Laird, P. W. (2002) ‘Hypomethylation and hypermethylation of DNA in Wilms tumors’, *Oncogene*, 21, pp. 6694–6702. doi: 10.1038/sj.onc.

Erger, F., Nörling, D., Borchert, D., Leenen, E., Habbig, S., Wiesener, M. S., Bartram, M. P., Wenzel, A., Becker, C., Toliat, M. R., Nürnberg, P., Beck, B. B. and Altmüller, J. (2020) ‘cfNOMe — A single assay for comprehensive epigenetic analyses of cell-free DNA’, *Genome Medicine*. Springer Science and Business Media LLC, 12(1), p. 54. doi:

10.1186/s13073-020-00750-5.

Esteller, M., Corn, P. G., Baylin, S. B. and Herman, J. G. (2001) 'A gene hypermethylation profile of human cancer', *Cancer Research*, 61(8), pp. 3225–3229.

Evans, J., Chapple, A., Salisbury, H., Corrie, P. and Ziebland, S. (2014) “‘It can’t be very important because it comes and goes’-patients’ accounts of intermittent symptoms preceding a pancreatic cancer diagnosis: A qualitative study’, *BMJ Open*. BMJ Publishing Group, 4(2). doi: 10.1136/bmjopen-2013-004215.

Fang, F., Turcan, S., Rimner, A., Kaufman, A., Giri, D., Morris, L. G. T., Shen, R., Seshan, V., Mo, Q., Heguy, A., Baylin, S. B., Ahuja, N., Viale, A., Massague, J., Norton, L., Vahdat, L. T., Moynahan, M. E. and Chan, T. A. (2011) 'Breast cancer methylomes establish an epigenomic foundation for metastasis', *Science Translational Medicine*. American Association for the Advancement of Science, 3(75), pp. 75ra25-75ra25. doi: 10.1126/scitranslmed.3001875.

Farrelly, L. A., Thompson, R. E., Zhao, S., Lepack, A. E., Lyu, Y., Bhanu, N. V., Zhang, B., Loh, Y.-H. E., Ramakrishnan, A., Vadodaria, K. C., Heard, K. J., Erikson, G., Nakadai, T., Bastle, R. M., Lukasak, B. J., Zebroski, H., Alenina, N., Bader, M., Berton, O., Roeder, R. G., Molina, H., Gage, F. H., Shen, L., Garcia, B. A., Li, H., Muir, T. W. and Maze, I. (2019) 'Histone serotonylation is a permissive modification that enhances TFIID binding to H3K4me3', *Nature*. Nature Publishing Group, 567(7749), pp. 535–539. doi: 10.1038/s41586-019-1024-7.

Fass, L. (2008) 'Imaging and cancer: A review', *Molecular Oncology*. No longer published by Elsevier, pp. 115–152. doi: 10.1016/j.molonc.2008.04.001.

Feng, Q. and Zhang, Y. (2001) 'The MeCP1 complex represses transcription through

preferential binding, remodeling, and deacetylating methylated nucleosomes’, *Genes and Development*, 15(7), pp. 827–832. doi: 10.1101/gad.876201.

Ficz, G., Branco, M. R., Seisenberger, S., Santos, F., Krueger, F., Hore, T. A., Marques, C. J., Andrews, S. and Reik, W. (2011) ‘Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation’, *Nature*. Nature Publishing Group, 473(7347), pp. 398–404. doi: 10.1038/nature10008.

Figuerola, M. E., Abdel-Wahab, O., Lu, C., Ward, P. S., Patel, J., Shih, A., Li, Y., Bhagwat, N., Vasanthakumar, A., Fernandez, H. F., Tallman, M. S., Sun, Z., Wolniak, K., Peeters, J. K., Liu, W., Choe, S. E., Fantin, V. R., Paietta, E., Löwenberg, B., Licht, J. D., Godley, L. A., Delwel, R., Valk, P. J. M., Thompson, C. B., Levine, R. L. and Melnick, A. (2010) ‘Leukemic IDH1 and IDH2 Mutations Result in a Hypermethylation Phenotype, Disrupt TET2 Function, and Impair Hematopoietic Differentiation’, *Cancer Cell*. Elsevier, 18(6), pp. 553–567. doi: 10.1016/j.ccr.2010.11.015.

Frangioni, J. V. (2008) ‘New technologies for human cancer imaging’, *Journal of Clinical Oncology*. American Society of Clinical Oncology, pp. 4012–4021. doi: 10.1200/JCO.2007.14.3065.

Frommer, M., McDonald, L. E., Millar, D. S., Collis, C. M., Watt, F., Grigg, G. W., Molloy, P. L. and Paul, C. L. (1992) ‘A genomic sequencing protocol that yields a positive display of 5- methylcytosine residues in individual DNA strands’, *Proceedings of the National Academy of Sciences of the United States of America*, 89(5), pp. 1827–1831. doi: 10.1073/pnas.89.5.1827.

Fu, H. Y., Wu, D. Sen, Zhou, H. R. and Shen, J. Z. (2014) ‘CpG island methylator phenotype and its relationship with prognosis in adult acute leukemia patients’,

Hematology. Maney Publishing, 19(6), pp. 329–337. doi: 10.1179/1607845413Y.00000000137.

Gama-sosa, M. A., Slagel, V. A., Trewyn, R. W., Oxenhandler, R., Kuo, K. C., Gehrke, C. W. and Ehrlich, M. (1983) ‘The 5-methylcytosine content of DNA from human tumors’, *Nucleic Acids Research*, 11(19), pp. 6883–6894. doi: 10.1093/nar/11.19.6883.

Gao, X. D., Qu, J. H., Chang, X. J., Lu, Y. Y., Bai, W. L., Wang, H., Xu, Z. X., An, L. J., Wang, C. P., Zeng, Z. and Yang, Y. P. (2014) ‘Hypomethylation of long interspersed nuclear element-1 promoter is associated with poor outcomes for curative resected hepatocellular carcinoma’, *Liver International*, 34(1), pp. 136–146. doi: 10.1111/liv.12264.

Gates, L. A., Foulds, C. E. and O’Malley, B. W. (2017) ‘Histone Marks in the “Driver’s Seat”: Functional Roles in Steering the Transcription Cycle’, *Trends in Biochemical Sciences*. Elsevier Ltd, 42(12), pp. 977–989. doi: 10.1016/j.tibs.2017.10.004.

Gkountela, S., Castro-Giner, F., Szczerba, B. M., Vetter, M., Landin, J., Scherrer, R., Krol, I., Scheidmann, M. C., Beisel, C., Stirnimann, C. U., Kurzeder, C., Heinzelmann-Schwarz, V., Rochlitz, C., Weber, W. P. and Aceto, N. (2019) ‘Circulating Tumor Cell Clustering Shapes DNA Methylation to Enable Metastasis Seeding’, *Cell*, January, pp. 98–112.e14. doi: 10.1016/j.cell.2018.11.046.

Goonetilleke, K. S. and Siriwardena, A. K. (2007) ‘Systematic review of carbohydrate antigen (CA 19-9) as a biochemical marker in the diagnosis of pancreatic cancer’, *European Journal of Surgical Oncology*. Elsevier, pp. 266–270. doi: 10.1016/j.ejso.2006.10.004.

Gore, A. J., Deitz, S. L., Palam, L. R., Craven, K. E. and Korc, M. (2014) ‘Pancreatic

cancer-associated retinoblastoma 1 dysfunction enables TGF- β to promote proliferation', *Journal of Clinical Investigation*. American Society for Clinical Investigation, 124(1), pp. 338–352. doi: 10.1172/JCI71526.

Graustein, A. D., Horne, D. J., Fong, J. J., Schwarz, F., Mefford, H. C., Peterson, G. J., Wells, R. D., Musvosvi, M., Shey, M., Hanekom, W. A., Hatherill, M., Scriba, T. J., Thuong, N. T. T., Mai, N. T. H., Caws, M., Bang, N. D., Dunstan, S. J., Thwaites, G. E., Varki, A., Angata, T. and Hawn, T. R. (2017) 'The SIGLEC14 null allele is associated with Mycobacterium tuberculosis- and BCG-induced clinical and immunologic outcomes', *Tuberculosis*. Churchill Livingstone, 104, pp. 38–45. doi: 10.1016/j.tube.2017.02.005.

Greger, V., Passarge, E., Hfpping, W. and Horsthemke, B. (1989) *Epigenetic changes may contribute to the formation and spontaneous regression of retinoblastoma*, *Hum Genet*.

Greten, F. R., Weber, C. K., Greten, T. F., Schneider, G., Wagner, M., Adler, G. and Schmid, R. M. (2002) 'Stat3 and NF- κ B activation prevents apoptosis in pancreatic carcinogenesis', *Gastroenterology*. W.B. Saunders, 123(6), pp. 2052–2063. doi: 10.1053/gast.2002.37075.

Greten, T. F., Papendorf, F., Bleck, J. S., Kirchhoff, T., Wohlberedt, T., Kubicka, S., Klempnauer, J., Galanski, M. and Manns, M. P. (2005) 'Survival rate in patients with hepatocellular carcinoma: A retrospective analysis of 389 patients', *British Journal of Cancer*. Nature Publishing Group, 92(10), pp. 1862–1868. doi: 10.1038/sj.bjc.6602590.

Gros, L., Ishchenko, A. A., Ide, H., Elder, R. H. and Saparbaev, M. K. (2004) 'The major human AP endonuclease (Ape1) is involved in the nucleotide incision repair pathway',

Nucleic Acids Research, 32(1), pp. 73–81. doi: 10.1093/nar/gkh165.

Gu, T. P., Guo, F., Yang, H., Wu, H. P., Xu, G. F., Liu, W., Xie, Z. G., Shi, L., He, X., Jin, S. G., Iqbal, K., Shi, Y. G., Deng, Z., Szabó, P. E., Pfeifer, G. P., Li, J. and Xu, G. L. (2011) ‘The role of Tet3 DNA dioxygenase in epigenetic reprogramming by oocytes’, *Nature*. Nature Publishing Group, 477(7366), pp. 606–612. doi: 10.1038/nature10443.

Guibert, N., Hu, Y., Feeney, N., Kuang, Y., Plagnol, V., Jones, G., Howarth, K., Beeler, J. F., Paweletz, C. P. and Oxnard, G. R. (2018) ‘Amplicon-based next-generation sequencing of plasma cell-free DNA for detection of driver and resistance mutations in advanced non-small cell lung cancer’, *Annals of Oncology*, 29(4), pp. 1049–1055. doi: 10.1093/annonc/mdy005.

Guler, G. D., Ning, Y., Ku, C. J., Phillips, T., McCarthy, E., Ellison, C. K., Bergamaschi, A., Collin, F., Lloyd, P., Scott, A., Antoine, M., Wang, W., Chau, K., Ashworth, A., Quake, S. R. and Levy, S. (2020) ‘Detection of early stage pancreatic cancer using 5-hydroxymethylcytosine signatures in circulating cell free DNA’, *Nature Communications*. Springer US, 11(1), pp. 1–12. doi: 10.1038/s41467-020-18965-w.

Guo, M., Li, X., Zhang, L., Liu, D., Du, W., Yin, D., Lyu, N., Zhao, G., Guo, C. and Tang, D. (2017) ‘Accurate quantification of 5-Methylcytosine, 5-Hydroxymethylcytosine, 5-Formylcytosine, and 5-Carboxylcytosine in genomic DNA from breast cancer by chemical derivatization coupled with ultra performance liquid chromatography- electrospray quadrupole time’, *Oncotarget*. Impact Journals LLC, 8(53), pp. 91248–91257. doi: 10.18632/oncotarget.20093.

Guo, M., Ren, J., Brock, M. V., Herman, J. G. and Carraway, H. E. (2008) ‘Promoter methylation of HIN-1 in the progression to esophageal squamous cancer.’, *Epigenetics* :

official journal of the DNA Methylation Society. *Epigenetics*, 3(6), pp. 336–341. doi: 10.4161/epi.3.6.7158.

Guo, S., Diep, D., Plongthongkum, N., Fung, H. L., Zhang, K. K. and Zhang, K. K. (2017) ‘Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tumor tissue-of-origin mapping from plasma DNA’, *Nature Genetics*. Nature Publishing Group, 49(4), pp. 635–642. doi: 10.1038/ng.3805.

Haab, B. B., Huang, Y., Balasenthil, S., Partyka, K., Tang, H., Anderson, M., Allen, P., Sasson, A., Zeh, H., Kaul, K., Kletter, D., Ge, S., Bern, M., Kwon, R., Blasutig, I., Srivastava, S., Frazier, M. L., Sen, S., Hollingsworth, M. A., Rinaudo, J. A., Killary, A. M. and Brand, R. E. (2015) ‘Definitive Characterization of CA 19-9 in Resectable Pancreatic Cancer Using a Reference Set of Serum and Plasma Specimens’, *PLOS ONE*. Edited by Z. Rakonczay. Public Library of Science, 10(10), p. e0139049. doi: 10.1371/journal.pone.0139049.

Hajkova, P., Jeffries, S. J., Lee, C., Miller, N., Jackson, S. P. and Surani, M. A. (2010) ‘Genome-wide reprogramming in the mouse germ line entails the base excision repair pathway’, *Science*. American Association for the Advancement of Science, 329(5987), pp. 78–82. doi: 10.1126/science.1187945.

Han, D. S. C., Ni, M., Chan, R. W. Y., Chan, V. W. H., Lui, K. O., Chiu, R. W. K. and Lo, Y. M. M. D. (2020) ‘The Biology of Cell-free DNA Fragmentation and the Roles of DNASE1, DNASE1L3, and DFFB’, *American Journal of Human Genetics*, 106(2), pp. 202–214. doi: 10.1016/j.ajhg.2020.01.008.

Han, H., Cortez, C. C., Yang, X., Nichols, P. W., Jones, P. A. and Liang, G. (2011) ‘DNA methylation directly silences genes with non-CpG island promoters and establishes a

nucleosome occupied promoter’, *Human Molecular Genetics*, 20(22), pp. 4299–4310. doi: 10.1093/hmg/ddr356.

Han, M., Jia, L., Lv, W., Wang, L. and Cui, W. (2019) ‘Epigenetic enzyme mutations: Role in tumorigenesis and molecular inhibitors’, *Frontiers in Oncology*. Frontiers, p. 194. doi: 10.3389/fonc.2019.00194.

Han, P., Fu, Y., Liu, Jingmei, Wang, Yunwu, He, J., Gong, J., Li, M., Tan, Q., Li, D., Luo, Y., Han, J., Liu, Jiqiao, Tu, W., Wang, Ying, Tian, D. and Yan, W. (2015) ‘Netrin-1 promotes cell migration and invasion by down-regulation of BVES expression in human hepatocellular carcinoma’, *American Journal of Cancer Research*. e-Century Publishing Corporation, 5(4), pp. 1396–1409. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/26101705> (Accessed: 23 August 2020).

Hanahan, D. and Weinberg, R. A. (2011) ‘Hallmarks of cancer: The next generation’, *Cell*. Elsevier, pp. 646–674. doi: 10.1016/j.cell.2011.02.013.

Hansen, K. D., Langmead, B. and Irizarry, R. A. (2012) ‘BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions’, *Genome Biology*. BioMed Central, 13(10), pp. 1–10. doi: 10.1186/gb-2012-13-10-R83.

Hansen, R. S., Stöger, R., Wijmenga, C., Stanek, A. M., Canfield, T. K., Luo, P., Matarazzo, M. R., D’Esposito, M., Feil, R., Gimelli, G., Weemaes, C. M. R., Laird, C. D. and Gartler, S. M. (2000) ‘Escape from gene silencing in ICF syndrome: Evidence for advanced replication time as a major determinant’, *Human Molecular Genetics*. Oxford University Press, 9(18), pp. 2575–2587. doi: 10.1093/hmg/9.18.2575.

Hashimoto, H., Pais, J. E., Dai, N., Corrêa, I. R., Zhang, X., Zheng, Y. and Cheng, X. (2015) ‘Structure of Naegleria Tet-like dioxygenase (NgTet1) in complexes with a

reaction intermediate 5-hydroxymethylcytosine DNA', *Nucleic Acids Research*, 43(22), pp. 10713–10721. doi: 10.1093/nar/gkv870.

Hashimoto, H., Pais, J. E., Zhang, X., Saleh, L., Fu, Z. Q., Dai, N., Corrêa, I. R., Zheng, Y. and Cheng, X. (2014) 'Structure of a Naegleria Tet-like dioxygenase in complex with 5-methylcytosine DNA', *Nature*, 506(7488), pp. 391–395. doi: 10.1038/nature12905.

Hata, K., Okano, M., Lei, H. and Li, E. (2002) 'Dnmt3L cooperates with the Dnmt3 family of de novo DNA methyltransferases to establish maternal imprints in mice', *Development*, 129(8), pp. 1983–1993.

He, Y.-F., Li, B.-Z., Li, Z., Liu, P., Wang, Y., Tang, Q., Ding, J., Jia, Y., Chen, Z., Li, L., Sun, Y., Li, X., Dai, Q., Song, C.-X., Zhang, K., He, C. and Xu, G.-L. (2011) 'Tet-Mediated Formation of 5-Carboxylcytosine and Its Excision by TDG in Mammalian DNA', *Science*, 333(6047), pp. 1303–1307. doi: 10.1126/science.1210944.

Heitzer, E., Haque, I. S., Roberts, C. E. S. and Speicher, M. R. (2018) 'Current and future perspectives of liquid biopsies in genomics-driven oncology', *Nature Reviews Genetics*, 8 February, pp. 71–88. doi: 10.1038/s41576-018-0071-5.

Henriksen, S. D., Madsen, P. H., Larsen, A. C., Johansen, M. B., Drewes, A. M., Pedersen, I. S., Krarup, H. and Thorlacius-Ussing, O. (2016) 'Cell-free DNA promoter hypermethylation in plasma as a diagnostic marker for pancreatic adenocarcinoma', *Clinical Epigenetics*. Springer Verlag, 8(1), p. 117. doi: 10.1186/s13148-016-0286-2.

Hermann, A., Goyal, R. and Jeltsch, A. (2004) 'The Dnmt1 DNA-(cytosine-C5)-methyltransferase methylates DNA processively with high preference for hemimethylated target sites', *Journal of Biological Chemistry*. American Society for Biochemistry and Molecular Biology, 279(46), pp. 48350–48359. doi:

10.1074/jbc.M403427200.

Hofer, A., Liu, Z. J. and Balasubramanian, S. (2019) ‘Detection, Structure and Function of Modified DNA Bases’, *Journal of the American Chemical Society*. American Chemical Society, 141(16), pp. 6420–6429. doi: 10.1021/jacs.9b01915.

Hore, T. A., Von Meyenn, F., Ravichandran, M., Bachman, M., Ficz, G., Oxley, D., Santos, F., Balasubramanian, S., Jurkowski, T. P. and Reik, W. (2016) ‘Retinol and ascorbate drive erasure of Epigenetic memory and enhance reprogramming to naïve pluripotency by complementary mechanisms’, *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 113(43), pp. 12202–12207. doi: 10.1073/pnas.1608679113.

Hotchkiss, R. D. (1948) ‘The quantitative separation of purines, pyrimidines, and nucleosides by paper chromatography’, *The Journal of biological chemistry*, 175(1), pp. 315–332. doi: 10.3891/acta.chem.scand.06-1030.

Hou, Y., Guo, H., Cao, C., Li, X., Hu, B., Zhu, P., Wu, X., Wen, L., Tang, F., Huang, Y. and Peng, J. (2016) ‘Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas’, *Cell Research*, 26(3), pp. 304–319. doi: 10.1038/cr.2016.23.

Hu, L., Li, Z., Cheng, J., Rao, Q., Gong, W., Liu, M., Shi, Y. G., Zhu, J., Wang, P. and Xu, Y. (2013) ‘Crystal Structure of TET2-DNA Complex: Insight into TET-Mediated 5mC Oxidation’, *Cell*, 155(7), pp. 1545–1555. doi: 10.1016/j.cell.2013.11.020.

Hu, L., Lu, J., Cheng, J., Rao, Q., Li, Z., Hou, H., Lou, Z., Zhang, L., Li, W., Gong, W., Liu, M., Sun, C., Yin, X., Li, J., Tan, X., Wang, P., Wang, Y., Fang, D., Cui, Q., Yang, P., He, C., Jiang, H., Luo, C. and Xu, Y. (2015) ‘Structural insight into substrate

preference for TET-mediated oxidation’, *Nature*, 527(7576), pp. 118–122. doi: 10.1038/nature15713.

Hu, S., Wan, J., Su, Y., Song, Q., Zeng, Y., Nguyen, H. N., Shin, J., Cox, E., Rho, H. S., Woodard, C., Xia, S., Liu, S., Lyu, H., Ming, G. L., Wade, H., Song, H., Qian, J. and Zhu, H. (2013) ‘DNA methylation presents distinct binding sites for human transcription factors’, *eLife*, 2013(2), pp. 1–16. doi: 10.7554/eLife.00726.

Hu, Y., Ulrich, B. C., Supplee, J., Kuang, Y., Lizotte, P. H., Feeney, N. B., Guibert, N. M., Awad, M. M., Wong, K. K., Janne, P. A., Paweletz, C. P. and Oxnard, G. R. (2018) ‘False-positive plasma genotyping due to clonal hematopoiesis’, *Clinical Cancer Research*. American Association for Cancer Research Inc., 24(18), pp. 4437–4443. doi: 10.1158/1078-0432.CCR-18-0143.

Huang, R., Grishagin, I., Wang, Y., Zhao, T., Greene, J., Obenauer, J. C., Ngan, D., Nguyen, D.-T., Guha, R., Jadhav, A., Southall, N., Simeonov, A. and Austin, C. P. (2019) ‘The NCATS BioPlanet – An Integrated Platform for Exploring the Universe of Cellular Signaling Pathways for Toxicology, Systems Biology, and Chemical Genomics’, *Frontiers in Pharmacology*. Frontiers Media S.A., 10(APR), p. 445. doi: 10.3389/fphar.2019.00445.

Huang, S. Z., Wei, M. N., Huang, J. R., Zhang, Z. J., Zhang, W. J., Jiang, Q. W., Yang, Y., Wang, H. Y., Jin, H. L., Wang, K., Xing, Z. H., Yuan, M. L., Li, Y., He, X. S., Shi, Z. and Zhou, Q. (2019) ‘Targeting TF-AKT/ERK-EGFR pathway suppresses the growth of hepatocellular carcinoma’, *Frontiers in Oncology*. Frontiers Media S.A., 9(MAR), p. 150. doi: 10.3389/fonc.2019.00150.

Hughes, L. A. E., Melotte, V., Schrijver, J. De, Maat, M. De, Smit, V. T. H. B. M., Bovee,

J. V. M. G., French, P. J., Van Den Brandt, P. A., Schouten, L. J., De Meyer, T., Van Criekinge, W., Ahuja, N., Herman, J. G., Weijenberg, M. P. and Van Engeland, M. (2013) ‘The CpG island methylator phenotype: What’s in a name?’, *Cancer Research*. American Association for Cancer Research, pp. 5858–5868. doi: 10.1158/0008-5472.CAN-12-4306.

Hur, C., Tramontano, A. C., Dowling, E. C., Brooks, G. A., Jeon, A., Brugge, W. R., Gazelle, G. S., Kong, C. Y. and Pandharipande, P. V. (2016) ‘Early Pancreatic Ductal Adenocarcinoma Survival Is Dependent on Size’, *Pancreas*, 45(7), pp. 1062–1066. doi: 10.1097/MPA.0000000000000587.

Imperiale, T. F., Ransohoff, D. F., Itzkowitz, S. H., Levin, T. R., Lavin, P., Lidgard, G. P., Ahlquist, D. A. and Berger, B. M. (2014) ‘Multitarget stool DNA testing for colorectal-cancer screening’, *New England Journal of Medicine*. Massachusetts Medical Society, 370(14), pp. 1287–1297. doi: 10.1056/NEJMoa1311194.

Incarnato, D., Krepelova, A. and Neri, F. (2014) ‘High-throughput single nucleotide variant discovery in E14 mouse embryonic stem cells provides a new reference genome assembly’, *Genomics*. Academic Press Inc., 104(2), pp. 121–127. doi: 10.1016/j.ygeno.2014.06.007.

Iqbal, K., Jin, S. G., Pfeifer, G. P. and Szabó, P. E. (2011) ‘Reprogramming of the paternal genome upon fertilization involves genome-wide oxidation of 5-methylcytosine’, *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 108(9), pp. 3642–3647. doi: 10.1073/pnas.1014033108.

Ishchenko, A. A., Sanz, G., Privezentzev, C. V., Maksimenko, A. V. and Saparbaev, M.

(2003) ‘Characterisation of new substrate specificities of *Escherichia coli* and *Saccharomyces cerevisiae* AP endonucleases’, *Nucleic Acids Research*, 31(21), pp. 6344–6353. doi: 10.1093/nar/gkg812.

Ito, S., Dalessio, A. C., Taranova, O. V., Hong, K., Sowers, L. C. and Zhang, Y. (2010) ‘Role of tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification’, *Nature*. Howard Hughes Medical Institute, 466(7310), pp. 1129–1133. doi: 10.1038/nature09303.

Ito, S., Shen, L., Dai, Q., Wu, S. C., Collins, L. B., Swenberg, J. A., He, C. and Zhang, Y. (2011) ‘Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine’, *Science*. American Association for the Advancement of Science, 333(6047), pp. 1300–1303. doi: 10.1126/science.1210597.

Iurlaro, M., Ficiz, G., Oxley, D., Raiber, E. A., Bachman, M., Booth, M. J., Andrews, S., Balasubramanian, S. and Reik, W. (2013) ‘A screen for hydroxymethylcytosine and formylcytosine binding proteins suggests functions in transcription and chromatin regulation’, *Genome Biology*. BioMed Central, 14(10), p. R119. doi: 10.1186/gb-2013-14-10-r119.

Iyer, L. M., Tahiliani, M., Rao, A. and Aravind, L. (2009) *Prediction of novel families of enzymes involved in oxidative and other complex modifications of bases in nucleic acids*, *Cell Cycle*.

Jain, M., Koren, S., Miga, K. H., Quick, J., Rand, A. C., Sasani, T. A., Tyson, J. R., Beggs, A. D., Dilthey, A. T., Fiddes, I. T., Malla, S., Marriott, H., Nieto, T., O’Grady, J., Olsen, H. E., Pedersen, B. S., Rhie, A., Richardson, H., Quinlan, A. R., Snutch, T. P., Tee, L., Paten, B., Phillippy, A. M., Simpson, J. T., Loman, N. J. and Loose, M. (2018)

‘Nanopore sequencing and assembly of a human genome with ultra-long reads’, *Nature Biotechnology*. Nature Publishing Group, 36(4), pp. 338–345. doi: 10.1038/nbt.4060.

Jakubowska, K., Guzińska-Ustymowicz, K., Famulski, W., Cepowicz, D., Jagodzińska, D. and Pryczynicz, A. (2016) ‘Reduced expression of caspase-8 and cleaved caspase-3 in pancreatic ductal adenocarcinoma cells’, *Oncology Letters*. Spandidos Publications, 11(3), pp. 1879–1884. doi: 10.3892/ol.2016.4125.

Jang, H. S., Shah, N. M., Du, A. Y., Dailey, Z. Z., Pehrsson, E. C., Godoy, P. M., Zhang, D., Li, D., Xing, X., Kim, S., O'Donnell, D., Gordon, J. I. and Wang, T. (2019) ‘Transposable elements drive widespread expression of oncogenes in human cancers’, *Nature Genetics*. Nature Publishing Group, pp. 611–617. doi: 10.1038/s41588-019-0373-3.

Jenuwein, T. and Allis, C. D. (2001) ‘Translating the histone code’, *Science*. American Association for the Advancement of Science, pp. 1074–1080. doi: 10.1126/science.1063127.

Jiang, D., Hatahet, Z., Melamede, R. J., Kow, Y. W. and Wallace, S. S. (1997) ‘1997_Jiang_Characterization of Escherichia coli Endonuclease VIII’, *Journal of Biological Chemistry*. American Society for Biochemistry and Molecular Biology, 272(51), pp. 32230–32239. doi: 10.1074/jbc.272.51.32230.

Jiang, P., Chan, C. W. M., Chan, K. C. A., Cheng, S. H., Wong, J., Wong, V. W.-S., Wong, G. L. H., Chan, S. L., Mok, T. S. K., Chan, H. L. Y., Lai, P. B. S., Chiu, R. W. K. and Lo, Y. M. D. (2015) ‘Lengthening and shortening of plasma DNA in hepatocellular carcinoma patients’, *Proceedings of the National Academy of Sciences*, 112(11), pp. E1317–E1325. doi: 10.1073/pnas.1500076112.

Jimmy Huang, P. C., Low, P. Y., Wang, I., Danny Hsu, S. Te and Angata, T. (2018) ‘Soluble Siglec-14 glycan-recognition protein is generated by alternative splicing and suppresses myeloid inflammatory responses’, *Journal of Biological Chemistry*. American Society for Biochemistry and Molecular Biology Inc., 293(51), pp. 19645–19658. doi: 10.1074/jbc.RA118.005676.

Johnson, T. B. and Coghill, R. D. (1925) ‘Researches on pyrimidines. C111. The discovery of 5-methyl-cytosine in tuberculinic acid, the nucleic acid of the tubercle bacillus’, *Journal of the American Chemical Society*. American Chemical Society, 47(11), pp. 2838–2844. doi: 10.1021/ja01688a030.

Kai, M., Niinuma, T., Kitajima, H., Yamamoto, E., Harada, T., Aoki, H., Maruyama, R., Toyota, M., Sasaki, Y., Sugai, T., Tokino, T., Nakase, H. and Suzuki, H. (2016) ‘TET1 Depletion Induces Aberrant CpG Methylation in Colorectal Cancer Cells’, *PLOS ONE*. Edited by J. S. Castresana. Public Library of Science, 11(12), p. e0168281. doi: 10.1371/journal.pone.0168281.

Kaneda, M., Okano, M., Hata, K., Sado, T., Tsujimoto, H., Li, E. and Sasaki, H. (2004) ‘Essential role for de novo DNA methyltransferase Dnmt3a in paternal and maternal imprinting’, *Nature*. Nature Publishing Group, 429(6994), pp. 900–903. doi: 10.1038/nature02633.

Kansagara, D., Papak, J., Pasha, A. S., O’Neil, M., Freeman, M., Relevo, R., Quiñones, A., Motu’apuaka, M. and Jou, J. H. (2014) ‘Screening for hepatocellular carcinoma in chronic liver disease: A systematic review’, *Annals of Internal Medicine*, 161(4), pp. 261–269. doi: 10.7326/M14-0558.

Kaseb, A. O., Sanchez, N. S., Sen, S., Kelley, R. K., Tan, B., Bocobo, A. G., Lim, K. H.,

Abdel-Wahab, R., Uemura, M., Pestana, R. C., Qiao, W., Xiao, L., Morris, J., Amin, H. M., Hassan, M. M., Rashid, A., Banks, K. C., Lanman, R. B., Talasaz, A. A., Mills-Shaw, K. R., George, B., Haque, A., Raghav, K. P. S., Wolff, R. A., Yao, J. C., Meric-Bernstam, F., Ikeda, S. and Kurzrock, R. (2019) 'Molecular profiling of hepatocellular carcinoma using circulating cell-free DNA', *Clinical Cancer Research*. American Association for Cancer Research, 25(20), pp. 6107–6118. doi: 10.1158/1078-0432.CCR-18-3341.

Kellinger, M. W., Song, C. X., Chong, J., Lu, X. Y., He, C. and Wang, D. (2012) '5-formylcytosine and 5-carboxylcytosine reduce the rate and substrate specificity of RNA polymerase II transcription', *Nature Structural and Molecular Biology*. NIH Public Access, 19(8), pp. 831–833. doi: 10.1038/nsmb.2346.

Keshet, I., Schlesinger, Y., Farkash, S., Rand, E., Hecht, M., Segal, E., Pikarski, E., Young, R. A., Niveleau, A., Cedar, H. and Simon, I. (2006) 'Evidence for an instructive mechanism of de novo methylation in cancer cells', *Nature Genetics*. Nature Publishing Group, pp. 149–153. doi: 10.1038/ng1719.

Kim, J., Bamlet, W. R., Oberg, A. L., Chaffee, K. G., Donahue, G., Cao, X. J., Chari, S., Garcia, B. A., Petersen, G. M. and Zaret, K. S. (2017) 'Detection of early pancreatic ductal adenocarcinoma with thrombospondin-2 & CA19-9 blood markers', *Science Translational Medicine*. American Association for the Advancement of Science, 9(398), p. eaah5583. doi: 10.1126/scitranslmed.aah5583.

Kim, J., Kollhoff, A., Bergmann, A. and Stubbs, L. (2003) 'Methylation-sensitive binding of transcription factor YY1 to an insulator sequence within the paternally expressed imprinted gene, Peg3', *Human Molecular Genetics*, 12(3), pp. 233–245. doi: 10.1093/hmg/ddg028.

Kim, M., Trinh, B. N., Long, T. I., Oghamian, S. and Laird, P. W. (2004) ‘Dnmt1 deficiency leads to enhanced microsatellite instability in mouse embryonic stem cells’, *Nucleic Acids Research*, 32(19), pp. 5742–5749. doi: 10.1093/nar/gkh912.

Kisiel, J. B., Dukek, B. A., V.S.R. Kanipakam, R., Ghaz, H. M., Yab, T. C., Berger, C. K., Taylor, W. R., Foote, P. H., Giama, N. H., Onyirioha, K., Abdallah, M. A., Burger, K. N., Slettedahl, S. W., Mahoney, D. W., Smyrk, T. C., Lewis, J. T., Giakoumopoulos, M., Allawi, H. T., Lidgard, G. P., Roberts, L. R. and Ahlquist, D. A. (2019) ‘Hepatocellular Carcinoma Detection by Plasma Methylated DNA: Discovery, Phase I Pilot, and Phase II Clinical Validation’, *Hepatology*. John Wiley and Sons Inc., 69(3), pp. 1180–1192. doi: 10.1002/hep.30244.

Klein, E. A., Hubbell, E., Maddala, T., Aravanis, A., Beausang, J. F., Filippova, D., Gross, S., Jamshidi, A., Kurtzman, K., Shen, L., Valouev, A., Venn, O., Zhang, N., Smith, D. A., Yeatman, T. J., Tibshirani, R., Williams, R. T., Hartman, A.-R., Seiden, M. and Liu, M. C. (2018) ‘Development of a comprehensive cell-free DNA (cfDNA) assay for early detection of multiple tumor types: The Circulating Cell-free Genome Atlas (CCGA) study.’, *Journal of Clinical Oncology*. American Society of Clinical Oncology, 36(15_suppl), p. 12021. doi: 10.1200/JCO.2018.36.15_suppl.12021.

Klemm, S. L., Shipony, Z. and Greenleaf, W. J. (2019) ‘Chromatin accessibility and the regulatory epigenome’, *Nature Reviews Genetics*, 23 April, pp. 207–220. doi: 10.1038/s41576-018-0089-8.

Ko, M., Huang, Y., Jankowska, A. M., Pape, U. J., Tahiliani, M., Bandukwala, H. S., An, J., Lamperti, E. D., Koh, K. P., Ganetzky, R., Liu, X. S., Aravind, L., Agarwal, S., Maciejewski, J. P. and Rao, A. (2010) ‘Impaired hydroxylation of 5-methylcytosine in myeloid cancers with mutant TET2’, *Nature*. Nature Publishing Group, 468(7325), pp.

839–843. doi: 10.1038/nature09586.

Koch, A., Joosten, S. C., Feng, Z., De Ruijter, T. C., Draht, M. X., Melotte, V., Smits, K. M., Veeck, J., Herman, J. G., Neste, L. Van, Criekinge, W. Van, De Meyer, T. and Engeland, M. Van (2018) ‘Analysis of DNA methylation in cancer: Location revisited’, *Nature Reviews Clinical Oncology*. doi: 10.1038/s41571-018-0004-4.

Koestler, D. C., Marsit, C. J., Christensen, B. C., Accomando, W., Langevin, S. M., Houseman, E. A., Nelson, H. H., Karagas, M. R., Wiencke, J. K. and Kelsey, K. T. (2012) ‘Peripheral blood immune cell methylation profiles are associated with nonhematopoietic cancers’, *Cancer Epidemiology Biomarkers and Prevention*, 21(8), pp. 1293–1302. doi: 10.1158/1055-9965.EPI-12-0361.

Kohli, R. M. and Zhang, Y. (2013) ‘TET enzymes, TDG and the dynamics of DNA demethylation’, *Nature*. Nature Publishing Group, pp. 472–479. doi: 10.1038/nature12750.

Kong, L., Tan, L., Lv, R., Shi, Z., Xiong, L., Wu, F., Rabidou, K., Smith, M., He, C., Zhang, L., Qian, Y., Ma, D., Lan, F., Shi, Y. and Shi, Y. G. (2016) ‘A primary role of TET proteins in establishment and maintenance of De Novo bivalency at CpG islands’, *Nucleic Acids Research*, 44(18), pp. 8682–8692. doi: 10.1093/nar/gkw529.

Kotoh, Y., Suehiro, Y., Saeki, I., Hoshida, T., Maeda, M., Iwamoto, T., Matsumoto, T., Hidaka, I., Ishikawa, T., Takami, T., Higaki, S., Fujii, I., Suzuki, C., Shindo, Y., Tokumitsu, Y., Nagano, H., Sakaida, I. and Yamasaki, T. (2020) ‘Novel Liquid Biopsy Test Based on a Sensitive Methylated SEPT9 Assay for Diagnosing Hepatocellular Carcinoma’, *Hepatology Communications*. Wiley, 4(3), pp. 461–470. doi: 10.1002/hep4.1469.

Kremer, D., Metzger, S., Kolb-Bachofen, V. and Kremer, D. (2012) ‘Quantitative measurement of genome-wide DNA methylation by a reliable and cost-efficient enzyme-linked immunosorbent assay technique’, *Analytical Biochemistry*. Academic Press, 422(2), pp. 74–78. doi: 10.1016/j.ab.2011.11.033.

Kriaucionis, S. and Heintz, N. (2009) ‘The nuclear DNA base 5-hydroxymethylcytosine is present in purkinje neurons and the brain’, *Science*, 324(5929), pp. 929–930. doi: 10.1126/science.1169786.

Kriukiene, E., Labrie, V., Khare, T., Urbanavičiute, G., Lapinaite, A., Koncėvičius, K., Li, D., Wang, T., Pai, S., Ptak, C., Gordevičius, J., Wang, S. C., Petronis, A. and Klimašauskas, S. (2013) ‘DNA unmethylome profiling by covalent capture of CpG sites’, *Nature Communications*. Nature Publishing Group, 4(1), pp. 1–10. doi: 10.1038/ncomms3190.

Krop, I. E., Sgroi, D., Porter, D. A., Lunetta, K. L., Le Vangie, R., Seth, P., Kaelin, C. M., Rhei, E., Bosenberg, M., Schnitt, S., Marks, J. R., Pagon, Z., Belina, D., Razumovic, J. and Polyak, K. (2001) ‘HIN-1, a putative cytokine highly expressed in normal but not cancerous mammary epithelial cells’, *Proceedings of the National Academy of Sciences of the United States of America*. Proc Natl Acad Sci U S A, 98(17), pp. 9796–9801. doi: 10.1073/pnas.171138398.

Krueger, F. and Andrews, S. R. (2011) ‘Bismark: A flexible aligner and methylation caller for Bisulfite-Seq applications’, *Bioinformatics*, 27(11), pp. 1571–1572. doi: 10.1093/bioinformatics/btr167.

Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S. L., Jagodnik, K. M., Lachmann, A., McDermott, M. G., Monteiro,

C. D., Gundersen, G. W. and Ma'ayan, A. (2016) 'Enrichr: a comprehensive gene set enrichment analysis web server 2016 update', *Nucleic acids research*. Nucleic Acids Res, 44(W1), pp. W90–W97. doi: 10.1093/nar/gkw377.

Kustanovich, A., Schwartz, R., Peretz, T. and Grinshpun, A. (2019) 'Life and death of circulating cell-free DNA', *Cancer Biology and Therapy*, pp. 1057–1067. doi: 10.1080/15384047.2019.1598759.

Lam, D. K., Dang, D., Flynn, A. N., Hardt, M. and Schmidt, B. L. (2015) 'TMPRSS2, a novel membrane-anchored mediator in cancer pain', *Pain*. Lippincott Williams and Wilkins, 156(5), pp. 923–930. doi: 10.1097/j.pain.0000000000000130.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., Fitzhugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., Levine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, Christina, Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chisoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A.,

Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Hong, M. L., Dubois, J., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., De La Bastide, M., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H. C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G. R., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kasprzyk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F. A., Stupka, E., Szustakowki, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S. P., Yeh, R. F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Raymond, Christopher, Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Patrinos, A. and Morgan, M. J. (2001) 'Initial sequencing and analysis of the human genome', *Nature*. Nature Publishing Group, 409(6822), pp.

860–921. doi: 10.1038/35057062.

Langmead, B. and Salzberg, S. L. (2012) ‘Fast gapped-read alignment with Bowtie 2’, *Nature Methods*. Nature Publishing Group, 9(4), pp. 357–359. doi: 10.1038/nmeth.1923.

Laskin, J., Jones, S., Aparicio, S., Chia, S., Ch’ng, C., Deyell, R., Eirew, P., Fok, A., Gelmon, K., Ho, C., Huntsman, D., Jones, M., Kasaian, K., Karsan, A., Leelakumari, S., Li, Y., Lim, H., Ma, Y., Mar, C., Martin, M., Moore, R., Mungall, A., Mungall, K., Pleasance, E., Rassekh, S. R., Renouf, D., Shen, Y., Schein, J., Schrader, K., Sun, S., Tinker, A., Zhao, E., Yip, S. and Marra, M. A. (2015) ‘Lessons learned from the application of whole-genome analysis to the treatment of patients with advanced cancers’, *Molecular Case Studies*, 1(1), p. a000570. doi: 10.1101/mcs.a000570.

Law, J. A. and Jacobsen, S. E. (2010) ‘Establishing, maintaining and modifying DNA methylation patterns in plants and animals’, *Nature Reviews Genetics*. Nature Publishing Group, pp. 204–220. doi: 10.1038/nrg2719.

Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E. and Storey, J. D. (2014) ‘sva: Surrogate Variable Analysis R package version 3.10.0’, [Http://Www.Bioconductor.Org/Packages/Release/Bioc/Html/Sva.Html](http://www.bioconductor.org/packages/release/bioc/html/sva.html).

Lehmann-Werman, R., Neiman, D., Zemmour, H., Moss, J., Magenheimer, J., Vaknin-Dembinsky, A., Rubertsson, S., Nellgård, B., Blennow, K., Zetterberg, H., Spalding, K., Haller, M. J., Wasserfall, C. H., Schatz, D. A., Greenbaum, C. J., Dorrell, C., Grompe, M., Zick, A., Hubert, A., Maoz, M., Fendrich, V., Bartsch, D. K., Golan, T., Ben Sasson, S. A., Zamir, G., Razin, A., Cedar, H., Shapiro, A. M. J., Glaser, B., Shemer, R., Dor, Y., Sasson, S. A. B., Zamir, G., Razin, A., Cedara, H., Shapiro, A. M. J., Glaser, B., Shemer, R. and Dor, Y. (2016) ‘Identification of tissue-specific cell death using methylation

patterns of circulating DNA', *Proceedings of the National Academy of Sciences of the United States of America*, 113(13), pp. E1826–E1834. doi: 10.1073/pnas.1519286113.

Lentini, A., Lagerwall, C., Vikingsson, S., Mjoseng, H. K., Douvlataniotis, K., Vogt, H., Green, H., Meehan, R. R., Benson, M. and Nestor, C. E. (2018) 'A reassessment of DNA-immunoprecipitation-based genomic profiling', *Nature Methods*. Nature Publishing Group, 15(7), pp. 499–504. doi: 10.1038/s41592-018-0038-7.

Leung, D., Ren, B., Coarfa, C., Ecker, J. R., Epstein, C. B., Jones, S. J. M. M., Eaton, M. L., Whitaker, J. W., Pfenning, A. R., Zhou, X., Onuchic, V., Costello, J. F., Kim, A.-R., Haussler, D., Hirst, M., Ziller, M. J., Boyer, L. A., Karlic, R., Chuah, E., Harris, R. A., Wang, J., Marra, M. A., Quon, G., Farh, K.-H. H., De Jager, P. L., Carles, A., Hong, C., Kulkarni, A., Bansal, M. S., Shores, N., Tsai, L.-H. H., Beaudet, A. E., Zhang, B., Zhang, Z., Stevens, M., Wu, J., Sarkar, A., McManus, M. T., Yen, A., Sinnott-Armstrong, N. A., Xie, W., Sallari, R. C., Heravi-Moussavi, A., Chadwick, L. H., Rajagopal, N., Schultz, M. D., Fisher, S. J., Ernst, J., Wu, Y.-C. C., Ward, L. D., Dixon, J. R., Kellis, M., Wang, X., Meissner, A., Li, D., Zhang, M. Q., Farnham, P. J., Polak, P., Bernstein, B. E., Sabo, P. J., Liu, Y., Milosavljevic, A., Meuleman, W., Waterland, R. A., Hawkins, R. D., Mungall, A. J., Li, W., Feizi, S., Gjoneska, E., Neph, S. J., Siebenthall, K. T., Mercer, T. R., Thomson, J. A., Claussnitzer, M., Moore, R., Tlsty, T. D., Kheradpour, P., Thurman, R. E., Gascard, P., Sandstrom, R. S., Wang, W., Bilenky, M., Kaul, R., Kundaje, A., Canfield, T. K., Stamatoyannopoulos, J. A., Tam, A., Wang, T., Lowdon, R., Lister, R., Elliott, G., Hansen, R. S., Ray, P., Sunyaev, S., Amin, V., Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., Amin, V., Whitaker, J. W., Schultz, M. D., Ward, L. D., Sarkar, A., Quon, G., Sandstrom, R. S.,

Eaton, M. L., Wu, Y.-C. C., Pfenning, A. R., Wang, X., Claussnitzer, M., Liu, Y., Coarfa, C., Harris, R. A., Shores, N., Epstein, C. B., Gjoneska, E., Leung, D., Xie, W., Hawkins, R. D., Lister, R., Hong, C., Gascard, P., Mungall, A. J., Moore, R., Chuah, E., Tam, A., Canfield, T. K., Hansen, R. S., Kaul, R., Sabo, P. J., Bansal, M. S., Carles, A., Dixon, J. R., Farh, K.-H. H., Feizi, S., Karlic, R., Kim, A.-R. R., Kulkarni, A., Li, D., Lowdon, R., Elliott, G., Mercer, T. R., Neph, S. J., Onuchic, V., Polak, P., Rajagopal, N., Ray, P., Sallari, R. C., Siebenthall, K. T., Sinnott-Armstrong, N. A., Stevens, M., Thurman, R. E., Wu, J., Zhang, B., Zhou, X., Beaudet, A. E., Boyer, L. A., De Jager, P. L., Farnham, P. J., Fisher, S. J., Haussler, D., Jones, S. J. M. M., Li, W., Marra, M. A., McManus, M. T., Sunyaev, S., Thomson, J. A., Tlsty, T. D., Tsai, L.-H. H., Wang, W., Waterland, R. A., Zhang, M. Q., Chadwick, L. H., Bernstein, B. E., Costello, J. F., Ecker, J. R., Hirst, M., Meissner, A., Milosavljevic, A., Ren, B., Stamatoyannopoulos, J. A., Wang, T. and Kellis, M. (2015) 'Integrative analysis of 111 reference human epigenomes', *Nature*. Nature Publishing Group, 518(7539), pp. 317–330. doi: 10.1038/nature14248.

Li, E., Beard, C. and Jaenisch, R. (1993) 'Role for DNA methylation in genomic imprinting', *Nature*. Nature Publishing Group, 366(6453), pp. 362–365. doi: 10.1038/366362a0.

Li, E., Bestor, T. H. and Jaenisch, R. (1992) 'Targeted mutation of the DNA methyltransferase gene results in embryonic lethality', *Cell*, 69(6), pp. 915–926. doi: 10.1016/0092-8674(92)90611-F.

Li, G., Xu, W., Zhang, L., Liu, T., Jin, G., Song, J., Wu, J., Wang, Y., Chen, W., Zhang, C., Chen, X., Ding, Z., Zhu, P. and Zhang, B. (2019) 'Development and validation of a CIMP-associated prognostic model for hepatocellular carcinoma', *EBioMedicine*. Elsevier B.V., 47, pp. 128–141. doi: 10.1016/j.ebiom.2019.08.064.

Li, H. and Durbin, R. (2010) ‘Fast and accurate long-read alignment with Burrows-Wheeler transform’, *Bioinformatics*, 26(5), pp. 589–595. doi: 10.1093/bioinformatics/btp698.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R. (2009) ‘The Sequence Alignment/Map format and SAMtools’, *Bioinformatics*, 25(16), pp. 2078–2079. doi: 10.1093/bioinformatics/btp352.

Li, J., Wang, W., Xia, P., Wan, L., Zhang, L., Yu, L., Wang, L., Chen, X., Xiao, Y. and Xu, C. (2018) ‘Identification of a five-lncRNA signature for predicting the risk of tumor recurrence in patients with breast cancer’, *International Journal of Cancer*. Wiley-Liss Inc., 143(9), pp. 2150–2160. doi: 10.1002/ijc.31573.

Li, W. and Liu, M. (2011) ‘Distribution of 5-hydroxymethylcytosine in different human tissues’, *Journal of Nucleic Acids*. Hindawi Limited, 2011. doi: 10.4061/2011/870726.

Li, X., Li, W. and Xu, Y. (2018) ‘Human age prediction based on DNA methylation using a gradient boosting regressor’, *Genes*. MDPI AG, 9(9). doi: 10.3390/genes9090424.

Lian, C. G., Xu, Y., Ceol, C., Wu, F., Larson, A., Dresser, K., Xu, W., Tan, L., Hu, Y., Zhan, Q., Lee, C. W., Hu, D., Lian, B. Q., Kleffel, S., Yang, Y., Neiswender, J., Khorasani, A. J., Fang, R., Lezcano, C., Duncan, L. M., Scolyer, R. A., Thompson, J. F., Kakavand, H., Houvras, Y., Zon, L. I., Mihm, M. C., Kaiser, U. B., Schatton, T., Woda, B. A., Murphy, G. F. and Shi, Y. G. (2012) ‘Loss of 5-hydroxymethylcytosine is an epigenetic hallmark of Melanoma’, *Cell*. Elsevier, 150(6), pp. 1135–1146. doi: 10.1016/j.cell.2012.07.033.

Lim, J. and Singal, A. G. (2019) ‘Surveillance and Diagnosis of Hepatocellular Carcinoma’, *Clinical Liver Disease*, 13(1), pp. 2–5. doi: 10.1002/cld.761.

Lin, C. H., Hsieh, S. Y., Sheen, I. S., Lee, W. C., Chen, T. C., Shyu, W. C. and Liaw, Y. F. (2001) 'Genome-wide hypomethylation in hepatocellular carcinogenesis', *Cancer Research*, 61(10), pp. 4238–4243.

Lister, R., Mukamel, E. A., Nery, J. R., Urich, M., Puddifoot, C. A., Johnson, N. D., Lucero, J., Huang, Y., Dwork, A. J., Schultz, M. D., Yu, M., Tonti-Filippini, J., Heyn, H., Hu, S., Wu, J. C., Rao, A., Esteller, M., He, C., Haghghi, F. G., Sejnowski, T. J., Behrens, M. M. and Ecker, J. R. (2013) 'Global epigenomic reconfiguration during mammalian brain development', *Science*. American Association for the Advancement of Science, 341(6146). doi: 10.1126/science.1237905.

Lister, R., Pelizzola, M., Dowen, R. H., Hawkins, R. D., Hon, G., Tonti-Filippini, J., Nery, J. R., Lee, L., Ye, Z., Ngo, Q. M., Edsall, L., Antosiewicz-Bourget, J., Stewart, R., Ruotti, V., Millar, A. H., Thomson, J. A., Ren, B. and Ecker, J. R. (2009) 'Human DNA methylomes at base resolution show widespread epigenomic differences', *Nature*. Nature Publishing Group, 462(7271), pp. 315–322. doi: 10.1038/nature08514.

Liu, D. S. K., Mato Prado, M., Giovannetti, E., Jiao, L. R., Krell, J. and Frampton, A. E. (2019) 'Can circulating tumor and exosomal nucleic acids act as biomarkers for pancreatic ductal adenocarcinoma?', *Expert Review of Molecular Diagnostics*. Taylor and Francis Ltd, 19(7), pp. 553–558. doi: 10.1080/14737159.2019.1622414.

Liu, J., Jiang, J., Mo, J., Liu, D., Cao, D., Wang, Hailin, He, Y. and Wang, Hongyang (2019) 'Global DNA 5-Hydroxymethylcytosine and 5-Formylcytosine Contents Are Decreased in the Early Stage of Hepatocellular Carcinoma', *Hepatology*. John Wiley and Sons Inc., 69(1), pp. 196–208. doi: 10.1002/hep.30146.

Liu, M. C., Oxnard, G. R., Klein, E. A., Swanton, C., Seiden, M. V, Liu, Minetta C.,

Oxnard, G. R., Klein, E. A., Smith, D., Richards, D., Yeatman, T. J., Cohn, A. L., Lapham, R., Clement, J., Parker, A. S., Tummala, M. K., McIntyre, K., Sekeres, M. A., Bryce, A. H., Siegel, R., Wang, X., Cosgrove, D. P., Abu-Rustum, N. R., Trent, J., Thiel, D. D., Becerra, C., Agrawal, M., Garbo, L. E., Giguere, J. K., Michels, R. M., Harris, R. P., Richey, S. L., McCarthy, T. A., Waterhouse, D. M., Couch, F. J., Wilks, S. T., Krie, A. K., Balaraman, R., Restrepo, A., Meshad, M. W., Rieger-Christ, K., Sullivan, T., Lee, C. M., Greenwald, D. R., Oh, W., Tsao, C. K., Fleshner, N., Kennecke, H. F., Khalil, M. F., Spigel, D. R., Manhas, A. P., Ulrich, B. K., Kover, P. A., Stokoe, C., Courtright, J. G., Yimer, H. A., Larson, T. G., Swanton, C., Seiden, M. V, Cummings, S. R., Absalan, F., Alexander, G., Allen, B., Amini, H., Aravanis, A. M., Bagaria, S., Bazargan, L., Beausang, J. F., Berman, J., Betts, C., Blocker, A., Bredno, J., Calef, R., Cann, G., Carter, J., Chang, C., Chawla, H., Chen, X., Chien, T. C., Civello, D., Davydov, K., Demas, V., Desai, M., Dong, Z., Fayzullina, S., Fields, A. P., Filippova, D., Freese, P., Fung, E. T., Gnerre, S., Gross, S., Halks-Miller, M., Hall, M. P., Hartman, A. R., Hou, C., Hubbell, E., Hunkapiller, N., Jagadeesh, K., Jamshidi, A., Jiang, R., Jung, B., Kim, T. H., Klausner, R. D., Kurtzman, K. N., Lee, M., Lin, W., Lipson, J., Liu, H., Liu, Q., Lopatin, M., Maddala, T., Maher, M. C., Melton, C., Mich, A., Nautiyal, S., Newman, Jonathan, Newman, Joshua, Nicula, V., Nicolaou, C., Nikolic, O., Pan, W., Patel, S., Prins, S. A., Rava, R., Ronaghi, N., Sakarya, O., Satya, R. V., Schellenberger, J., Scott, E., Sehnert, A. J., Shaknovich, R., Shanmugam, A., Shashidhar, K. C., Shen, L., Shenoy, A., Shojaei, S., Singh, P., Steffen, K. K., Tang, S., Tounge, J. M., Valouev, A., Venn, O., Williams, R. T., Wu, T., Xu, H. H., Yakym, C., Yang, X., Yecies, J., Yip, A. S., Youngren, J., Yue, J., Zhang, J., Zhang, L., Zhang, L. (Quan), Zhang, N., Curtis, C. and Berry, D. A. (2020) 'Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA', *Annals of Oncology*, 31(6), pp. 745–759. doi:

10.1016/j.annonc.2020.02.011.

Liu, M. Y., Denizio, J. E. and Kohli, R. M. (2016) ‘Quantification of Oxidized 5-Methylcytosine Bases and TET Enzyme Activity’, in *Methods in Enzymology*, pp. 365–385. doi: 10.1016/bs.mie.2015.12.006.

Liu, M. Y., Torabifard, H., Crawford, daniel J., DeNizio, J. E., Cao, X. J., Garcia, B. A., Cisneros, G. A. and Kohli, rahul M. (2017) ‘Mutations along a TET2 active site scaffold stall oxidation at 5-hydroxymethylcytosine’, *Nature Chemical Biology*, 13(2), pp. 181–187. doi: 10.1038/nchembio.2250.

Liu, Y., Cheng, J., Siejka-Zielińska, P., Weldon, C., Roberts, H., Lopopolo, M., Magri, A., D’Arienzo, V., Harris, J. M., McKeating, J. A. and Song, C. X. (2020) ‘Accurate targeted long-read DNA methylation and hydroxymethylation sequencing with TAPS’, *Genome Biology*. BioMed Central, 21(1), p. 54. doi: 10.1186/s13059-020-01969-6.

Liu, Y., Siejka-Zielińska, P., Velikova, G., Bi, Y., Yuan, F., Tomkova, M., Bai, C., Chen, L., Schuster-Böckler, B. and Song, C.-X. (2019) ‘Bisulfite-free direct detection of 5-methylcytosine and 5-hydroxymethylcytosine at base resolution’, *Nature Biotechnology*, 37(4), pp. 424–429. doi: 10.1038/s41587-019-0041-2.

Lomberk, G., Blum, Y., Nicolle, R., Nair, A., Gaonkar, K. S., Marisa, L., Mathison, A., Sun, Z., Yan, H., Elarouci, N., Armenoult, L., Ayadi, M., Ordog, T., Lee, J. H., Oliver, G., Klee, E., Moutardier, V., Gayet, O., Bian, B., Duconseil, P., Gilabert, M., Bigonnet, M., Garcia, S., Turrini, O., Delpero, J. R., Giovannini, M., Grandval, P., Gasmi, M., Secq, V., De Reyniès, A., Duseti, N., Iovanna, J. and Urrutia, R. (2018) ‘Distinct epigenetic landscapes underlie the pathobiology of pancreatic cancer subtypes’, *Nature Communications*, 9(1). doi: 10.1038/s41467-018-04383-6.

Luger, K., Mäder, A. W., Richmond, R. K., Sargent, D. F. and Richmond, T. J. (1997) 'Crystal structure of the nucleosome core particle at 2.8 Å resolution', *Nature*. Nature Publishing Group, 389(6648), pp. 251–260. doi: 10.1038/38444.

Majumder, S., Raimondo, M., Taylor, W. R., Yab, T. C., Berger, C. K., Dukek, B. A., Cao, X., Foote, P. H., Wu, C. W., Devens, M. E., Mahoney, D. W., Smyrk, T. C., Pannala, R., Chari, S. T., Vege, S. S., Topazian, M. D., Petersen, B. T., Levy, M. J., Rajan, E., Gleeson, F. C., Dayyeh, B. A., Nguyen, C. C., Faigel, D. O., Woodward, T. A., Wallace, M. B., Petersen, G., Allawi, H. T., Lidgard, G. P., Kisiel, J. B. and Ahlquist, D. A. (2019) 'Methylated DNA in Pancreatic Juice Distinguishes Patients With Pancreatic Cancer From Controls', *Clinical Gastroenterology and Hepatology*. W.B. Saunders. doi: 10.1016/j.cgh.2019.07.017.

Malnic, B., Godfrey, P. A. and Buck, L. B. (2004) 'The human olfactory receptor gene family', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 101(8), pp. 2584–2589. doi: 10.1073/pnas.0307882100.

Mannelli, C. (2019) 'Tissue vs Liquid Biopsies for Cancer Detection: Ethical Issues', *Journal of Bioethical Inquiry*. Springer, 16(4), pp. 551–557. doi: 10.1007/s11673-019-09944-y.

Martin-Herranz, D. E., Ribeiro, A. J. M., Krueger, F., Thornton, J. M., Reik, W. and Stubbs, T. M. (2017) 'cuRRBS: simple and robust evaluation of enzyme combinations for reduced representation approaches', *Nucleic Acids Research*, 45(20), pp. 11559–11569. doi: 10.1093/nar/gkx814.

Martinez-Arguelles, D. B., Lee, S. and Papadopoulos, V. (2014) 'In silico analysis

identifies novel restriction enzyme combinations that expand reduced representation bisulfite sequencing CpG coverage’, *BMC Research Notes*. BioMed Central Ltd., 7(1), p. 534. doi: 10.1186/1756-0500-7-534.

Matsuzaki, H., Okamura, E., Kuramochi, D., Ushiki, A., Hirakawa, K., Fukamizu, A. and Tanimoto, K. (2018) ‘Synthetic DNA fragments bearing ICR cis elements become differentially methylated and recapitulate genomic imprinting in transgenic mice’, *Epigenetics and Chromatin*. BioMed Central, 11(1), pp. 1–11. doi: 10.1186/s13072-018-0207-z.

De Mattos-Arruda, L., Mayor, R., Ng, C. K. Y., Weigelt, B., Martínez-Ricarte, F., Torrejon, D., Oliveira, M., Arias, A., Raventos, C., Tang, J., Guerini-Rocco, E., Martínez-Saéz, E., Lois, S., Marín, O., De La Cruz, X., Piscuoglio, S., Towers, R., Vivancos, A., Peg, V., Cajal, S. R. Y., Carles, J., Rodon, J., González-Cao, M., Tabernero, J., Felip, E., Sahuquillo, J., Berger, M. F., Cortes, J., Reis-Filho, J. S. and Seoane, J. (2015) ‘Cerebrospinal fluid-derived circulating tumour DNA better represents the genomic alterations of brain tumours than plasma’, *Nature Communications*. Nature Publishing Group, 6. doi: 10.1038/ncomms9839.

Maunakea, A. K., Nagarajan, R. P., Bilenky, M., Ballinger, T. J., Dsouza, C., Fouse, S. D., Johnson, B. E., Hong, C., Nielsen, C., Zhao, Y., Turecki, G., Delaney, A., Varhol, R., Thiessen, N., Shchors, K., Heine, V. M., Rowitch, D. H., Xing, X., Fiore, C., Schillebeeckx, M., Jones, S. J. M., Haussler, D., Marra, M. A., Hirst, M., Wang, T. and Costello, J. F. (2010) ‘Conserved role of intragenic DNA methylation in regulating alternative promoters’, *Nature*. Nature Publishing Group, 466(7303), pp. 253–257. doi: 10.1038/nature09165.

Maxim, L. D., Niebo, R. and Utell, M. J. (2014) ‘Screening tests: A review with

examples', *Inhalation Toxicology*. Informa Healthcare, pp. 811–828. doi: 10.3109/08958378.2014.955932.

Mayer, W., Niveleau, A., Walter, J., Fundele, R. and Haaf, T. (2000) 'Demethylation of the zygotic paternal genome', *Nature*. Nature Publishing Group, 403(6769), pp. 501–502. doi: 10.1038/35000656.

McDonald, O. G., Li, X., Saunders, T., Tryggvadottir, R., Mentch, S. J., Warmoes, M. O., Word, A. E., Carrer, A., Salz, T. H., Natsume, S., Stauffer, K. M., Makohon-Moore, A., Zhong, Y., Wu, H., Wellen, K. E., Locasale, J. W., Iacobuzio-Donahue, C. A. and Feinberg, A. P. (2017) 'Epigenomic reprogramming during pancreatic cancer progression links anabolic glucose metabolism to distant metastasis', *Nature Genetics*. Nature Publishing Group, 49(3), pp. 367–376. doi: 10.1038/ng.3753.

McPhail, S., Johnson, S., Greenberg, D., Peake, M. and Rous, B. (2015) 'Stage at diagnosis and early mortality from cancer in England', *British Journal of Cancer*. Nature Publishing Group, 112(Suppl 1), p. S108. doi: 10.1038/bjc.2015.49.

Meissner, A., Gnirke, A., Bell, G. W., Ramsahoye, B., Lander, E. S. and Jaenisch, R. (2005) 'Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis', *Nucleic Acids Research*, 33(18), pp. 5868–5877. doi: 10.1093/nar/gki901.

Mohapatra, S. S. and Biondi, E. G. (2017) 'DNA Methylation in Prokaryotes: Regulation and Function', in *Cellular Ecophysiology of Microbe*. Springer International Publishing, pp. 1–21. doi: 10.1007/978-3-319-20796-4_23-1.

Monnot, S., Serre, V., Chadeaux-Vekemans, B., Aupetit, J., Romano, S., De Lonlay, P., Rival, J. M., Munnich, A., Steffann, J. and Bonnefont, J. P. (2009) 'Structural insights on

pathogenic effects of novel mutations causing pyruvate carboxylase deficiency’, *Human Mutation*. Hum Mutat, 30(5), pp. 734–740. doi: 10.1002/humu.20908.

Moss, J., Magenheimer, J., Neiman, D., Korach, A., Samet, Y., Maoz, M., Zick, A., Grinshpun, A., Shapiro, A. M. J. J., Grompe, M., Spalding, K. L., Wittenberg, A. D., Glaser, B., Shemer, R., Zemmour, H., Loyfer, N., Korach, A., Samet, Y., Maoz, M., Druid, H., Arner, P., Fu, K. Y., Kiss, E., Spalding, K. L., Landesberg, G., Zick, A., Grinshpun, A., Shapiro, A. M. J. J., Grompe, M., Wittenberg, A. D., Glaser, B., Shemer, R., Kaplan, T. and Dor, Y. (2018) ‘Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease’, *Nature Communications*. Springer US, 9(1). doi: 10.1101/448142.

Mouliere, F., Chandrananda, D., Piskorz, A. M., Moore, E. K., Morris, J., Ahlborn, L. B., Mair, R., Goranova, T., Marass, F., Heider, K., Wan, J. C. M. M., Supernat, A., Hudecova, I., Gounaris, I., Ros, S., Jimenez-Linan, M., Garcia-Corbacho, J., Patel, K., Østrup, O., Murphy, S., Eldridge, M. D., Gale, D., Stewart, G. D., Burge, J., Cooper, W. N., Van Der Heijden, M. S., Massie, C. E., Watts, C., Corrie, P., Pacey, S., Brindle, K. M., Baird, R. D., Mau-Sørensen, M., Parkinson, C. A., Smith, C. G., Brenton, J. D., Rosenfeld, N., Heijden, M. S. Van Der, Massie, C. E., Watts, C., Corrie, P., Pacey, S., Brindle, K. M., Baird, R. D. and Mau-Sørensen, M. (2018) ‘Enhanced detection of circulating tumor DNA by fragment size analysis’, *Sci. Transl. Med*, 4921(November), p. 4921. doi: 10.1126/scitranslmed.aat4921.

Nair, S. S., Coolen, M. W., Stirzaker, C., Song, J. Z., Statham, A. L., Strbenac, D., Robinson, M. D. and Clark, S. J. (2011) ‘Comparison of methyl-DNA immunoprecipitation (MeDIP) and methyl-CpG binding domain (MBD) protein capture for genome-wide DNA methylation analysis reveal CpG sequence coverage bias’,

Epigenetics, 6(1), pp. 34–44. doi: 10.4161/epi.6.1.13313.

Nassiri, F., Chakravarthy, A., Feng, S., Shen, S. Y., Nejad, R., Zuccato, J. A., Voisin, M. R., Patil, V., Horbinski, C., Aldape, K., Zadeh, G. and De Carvalho, D. D. (2020) ‘Detection and discrimination of intracranial tumors using plasma cell-free DNA methylomes’, *Nature Medicine*. Springer US, 26(7), pp. 1044–1047. doi: 10.1038/s41591-020-0932-2.

Natale, F., Vivo, M., Falco, G. and Angrisano, T. (2019) ‘Deciphering DNA methylation signatures of pancreatic cancer and pancreatitis’, *Clinical Epigenetics*. BioMed Central, p. 132. doi: 10.1186/s13148-019-0728-8.

Navas, C., Hernández-Porras, I., Schuhmacher, A. J., Sibilia, M., Guerra, C. and Barbacid, M. (2012) ‘EGF Receptor Signaling Is Essential for K-Ras Oncogene-Driven Pancreatic Ductal Adenocarcinoma’, *Cancer Cell*. Europe PMC Funders, 22(3), pp. 318–330. doi: 10.1016/j.ccr.2012.08.001.

Neal, R. D., Tharmanathan, P., France, B., Din, N. U., Cotton, S., Fallon-Ferguson, J., Hamilton, W., Hendry, A., Hendry, M., Lewis, R., Macleod, U., Mitchell, E. D., Pickett, M., Rai, T., Shaw, K., Stuart, N., Tørring, M. L., Wilkinson, C., Williams, B., Williams, N. and Emery, J. (2015) ‘Is increased time to diagnosis and treatment in symptomatic cancer associated with poorer outcomes? Systematic review’, *British Journal of Cancer*, 112, pp. S92–S107. doi: 10.1038/bjc.2015.48.

Neoptolemos, J. P., Urrutia, R., Abbruzzese, J. L. and Büchler, M. W. (2018) *Pancreatic Cancer, Pancreatic Cancer*. Edited by J. P. Neoptolemos, R. Urrutia, J. L. Abbruzzese, and M. W. Büchler. New York, NY: Springer New York. doi: 10.1007/978-1-4939-7193-0.

Neri, F., Rapelli, S., Krepelova, A., Incarnato, D., Parlato, C., Basile, G., Maldotti, M., Anselmi, F. and Oliviero, S. (2017) 'Intragenic DNA methylation prevents spurious transcription initiation', *Nature*. Nature Publishing Group, 543(7643), pp. 72–77. doi: 10.1038/nature21373.

Nesterova, T. B., Popova, B. C., Cobb, B. S., Norton, S., Senner, C. E., Tang, Y. A., Spruce, T., Rodriguez, T. A., Sado, T., Merkenschlager, M. and Brockdorff, N. (2008) 'Dicer regulates Xist promoter methylation in ES cells indirectly through transcriptional control of Dnmt3a', *Epigenetics & Chromatin*. Springer Nature, 1(1), p. 2. doi: 10.1186/1756-8935-1-2.

Ngamruengphong, S. and Canto, M. I. (2016) 'Screening for Pancreatic Cancer', *Surgical Clinics of North America*, pp. 1223–1233. doi: 10.1016/j.suc.2016.07.016.

Nicolle, R., Blum, Y., Marisa, L., Loncle, C., Gayet, O., Moutardier, V., Turrini, O., Giovannini, M., Bian, B., Bigonnet, M., Rubis, M., Elarouci, N., Armenoult, L., Ayadi, M., Duconseil, P., Gasmi, M., Ouaisi, M., Maignan, A., Lomberk, G., Boher, J. M., Ewald, J., Bories, E., Garnier, J., Goncalves, A., Poizat, F., Raoul, J. L., Secq, V., Garcia, S., Grandval, P., Barraud-Blanc, M., Norguet, E., Gilabert, M., Delpero, J. R., Roques, J., Calvo, E., Guillaumond, F., Vasseur, S., Urrutia, R., de Reyniès, A., Dusetti, N. and Iovanna, J. (2017) 'Pancreatic Adenocarcinoma Therapeutic Targets Revealed by Tumor-Stroma Cross-Talk Analyses in Patient-Derived Xenografts', *Cell Reports*. Elsevier B.V., 21(9), pp. 2458–2470. doi: 10.1016/j.celrep.2017.11.003.

Nones, K., Waddell, N., Song, S., Patch, A. M., Miller, D., Johns, A., Wu, J., Kassahn, K. S., Wood, D., Bailey, P., Fink, L., Manning, S., Christ, A. N., Nourse, C., Kazakoff, S., Taylor, D., Leonard, C., Chang, D. K., Jones, M. D., Thomas, M., Watson, C., Pinese, M., Cowley, M., Rooman, I., Pajic, M., Butturini, G., Malpaga, A., Corbo, V., Crippa,

S., Falconi, M., Zamboni, G., Castelli, P., Lawlor, R. T., Gill, A. J., Scarpa, A., Pearson, J. V., Biankin, A. V. and Grimmond, S. M. (2014) 'Genome-wide DNA methylation patterns in pancreatic ductal adenocarcinoma reveal epigenetic deregulation of SLIT-ROBO, ITGA2 and MET signaling', *International Journal of Cancer*, 135(5), pp. 1110–1118. doi: 10.1002/ijc.28765.

Nosho, K., Shima, K., Irahara, N., Kure, S., Baba, Y., Kirkner, G. J., Chen, L., Gokhale, S., Hazra, A., Spiegelman, D., Giovannucci, E. L., Jaenisch, R., Fuchs, C. S. and Ogino, S. (2009) 'DNMT3B expression might contribute to CpG island methylator phenotype in colorectal cancer', *Clinical Cancer Research*. NIH Public Access, 15(11), pp. 3663–3671. doi: 10.1158/1078-0432.CCR-08-2383.

Nuzzo, P. V., Berchuck, J. E., Korthauer, K., Spisak, S., Nassar, A. H., Abou Alaiwi, S., Chakravarthy, A., Shen, S. Y., Bakouny, Z., Boccardo, F., Steinharter, J., Bouchard, G., Curran, C. R., Pan, W., Baca, S. C., Seo, J.-H., Lee, G.-S. M., Michaelson, M. D., Chang, S. L., Waikar, S. S., Sonpavde, G., Irizarry, R. A., Pomerantz, M., De Carvalho, D. D., Choueiri, T. K. and Freedman, M. L. (2020) 'Detection of renal cell carcinoma using plasma and urine cell-free DNA methylomes', *Nature Medicine*. Springer US, 26(7), pp. 1041–1043. doi: 10.1038/s41591-020-0933-1.

Okano, M., Bell, D. W., Haber, D. A. and Li, E. (1999) 'DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development', *Cell*. Cell Press, 99(3), pp. 247–257. doi: 10.1016/S0092-8674(00)81656-6.

Okano, M., Xie, S. and Li, E. (1998) 'Cloning and characterization of a family of novel mammalian DNA (cytosine-5) methyltransferases [1]', *Nature Genetics*. Nature Publishing Group, pp. 219–220. doi: 10.1038/890.

Okitsu, C. Y., Hsieh, J. C. F. and Hsieh, C.-L. (2010) 'Transcriptional Activity Affects the H3K4me3 Level and Distribution in the Coding Region', *Molecular and Cellular Biology*. American Society for Microbiology, 30(12), pp. 2933–2946. doi: 10.1128/mcb.01478-09.

Olova, N., Krueger, F., Andrews, S., Oxley, D., Berrens, R. V., Branco, M. R. and Reik, W. (2018) 'Comparison of whole-genome bisulfite sequencing library preparation strategies identifies sources of biases affecting DNA methylation data', *Genome Biology*, 19(1), p. 33. doi: 10.1186/s13059-018-1408-2.

Olsson, E., Winter, C., George, A., Chen, Y., Howlin, J., Tang, M. E., Dahlgren, M., Schulz, R., Grabau, D., Westen, D., Fernö, M., Ingvar, C., Rose, C., Bendahl, P., Rydén, L., Borg, Å., Gruvberger-Saal, S. K., Jernström, H. and Saal, L. H. (2015) 'Serial monitoring of circulating tumor DNA in patients with primary breast cancer for detection of occult metastatic disease', *EMBO Molecular Medicine*. EMBO, 7(8), pp. 1034–1047. doi: 10.15252/emmm.201404913.

Oxnard, G. R., Klein, E. A., Seiden, M. V., Hubbell, E., Venn, O., Jamshidi, A., Zhang, N., Beausang, J. F., Gross, S., Kurtzman, K. N., Fung, E. T., Allen, B., Fields, A. P., Liu, H., Sekeres, M. A., Richards, D. A., Yu, P. P., Aravanis, A. M., Hartman, A.-R. and Liu, M. C. (2019) 'Simultaneous multi-cancer detection and tissue of origin (TOO) localization using targeted bisulfite sequencing of plasma cell-free DNA (cfDNA)', *Annals of Oncology*, 30(October), p. v912. doi: 10.1093/annonc/mdz394.074.

Pais, J. E., Dai, N., Tamanaha, E., Vaisvila, R., Fomenkov, A. I., Bitinaite, J., Sun, Z., Guan, S., Corrêa, I. R., Noren, C. J., Cheng, X., Roberts, R. J., Zheng, Y., Saleh, L. and Rao, A. (2015) 'Biochemical characterization of a Naegleria TET-like oxygenase and its application in single molecule sequencing of 5-methylcytosine', *Proceedings of the*

National Academy of Sciences of the United States of America, 112(14), pp. 4316–4321.
doi: 10.1073/pnas.1417939112.

Parikh, A. R., Leshchiner, I., Elagina, L., Goyal, L., Levovitz, C., Siravegna, G., Livitz, D., Rhrissorakkrai, K., Martin, E. E., Van Seventer, E. E., Hanna, M., Slowik, K., Utro, F., Pinto, C. J., Wong, A., Danysh, B. P., de la Cruz, F. F., Fetter, I. J., Nadres, B., Shahzade, H. A., Allen, J. N., Blaszkowsky, L. S., Clark, J. W., Giantonio, B., Murphy, J. E., Nipp, R. D., Roeland, E., Ryan, D. P., Weekes, C. D., Kwak, E. L., Faris, J. E., Wo, J. Y., Aguet, F., Dey-Guha, I., Hazar-Rethinam, M., Dias-Santagata, D., Ting, D. T., Zhu, A. X., Hong, T. S., Golub, T. R., Iafrate, A. J., Adalsteinsson, V. A., Bardelli, A., Parida, L., Juric, D., Getz, G. and Corcoran, R. B. (2019) ‘Liquid versus tissue biopsy for detecting acquired resistance and tumor heterogeneity in gastrointestinal cancers’, *Nature Medicine*. Nature Publishing Group, 25(9), pp. 1415–1421. doi: 10.1038/s41591-019-0561-9.

Parikh, S. S., Walcher, G., Jones, G. D., Slupphaug, G., Krokan, H. E., Blackburn, G. M. and Tainer, J. A. (2000) ‘Uracil-DNA glycosylase-DNA substrate and product structures: Conformational strain promotes catalytic efficiency by coupled stereoelectronic effects’, *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 97(10), pp. 5083–5088. doi: 10.1073/pnas.97.10.5083.

Pascual, S., Miralles, C., Bernabé, J. M., Irurzun, J. and Planells, M. (2019) ‘Surveillance and diagnosis of hepatocellular carcinoma: A systematic review’, *World Journal of Clinical Cases*, 7(16), pp. 2269–2286. doi: 10.12998/wjcc.v7.i16.2269.

Peng, J., Xia, B. and Yi, C. (2016) ‘Single-base resolution analysis of DNA epigenome via high-throughput sequencing’, *Science China Life Sciences*, pp. 219–226. doi: 10.1007/s11427-016-5013-x.

Peter, M. E., Hadji, A., Murmann, A. E., Brockway, S., Putzbach, W., Pattanayak, A. and Ceppi, P. (2015) 'The role of CD95 and CD95 ligand in cancer', *Cell Death and Differentiation*. Nature Publishing Group, pp. 549–559. doi: 10.1038/cdd.2015.3.

Phallen, J., Sausen, M., Adleff, V., Leal, A., Hruban, C., White, J., Anagnostou, V., Fiksel, J., Cristiano, S., Papp, E., Speir, S., Reinert, T., Orntoft, M. B. W., Woodward, B. D., Murphy, D., Parpart-Li, S., Riley, D., Nesselbush, M., Sengamalay, N., Georgiadis, A., Li, Q. K., Madsen, M. R., Mortensen, F. V., Huiskens, J., Punt, C., Van Grieken, N., Fijneman, R., Meijer, G., Husain, H., Scharpf, R. B., Diaz, L. A., Jones, S., Angiuoli, S., Ørntoft, T., Nielsen, H. J., Andersen, C. L. and Velculescu, V. E. (2017) 'Direct detection of early-stage cancers using circulating tumor DNA', *Science Translational Medicine*. American Association for the Advancement of Science, 9(403). doi: 10.1126/scitranslmed.aan2415.

Pickhardt, P. J. (2016) 'Emerging stool-based and blood-based non-invasive DNA tests for colorectal cancer screening: the importance of cancer prevention in addition to cancer detection', *Abdominal Radiology*. Springer New York LLC, 41(8), pp. 1441–1444. doi: 10.1007/s00261-016-0798-4.

Pidsley, R., Zotenko, E., Peters, T. J., Lawrence, M. G., Risbridger, G. P., Molloy, P., Van Dijk, S., Muhlhausler, B., Stirzaker, C. and Clark, S. J. (2016) 'Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling', *Genome Biology*. Genome Biology, 17(1), pp. 1–17. doi: 10.1186/s13059-016-1066-1.

Plongthongkum, N., Diep, D. H. and Zhang, K. (2014) *Advances in the profiling of DNA modifications: Cytosine methylation and beyond*, *Nature Reviews Genetics*. doi: 10.1038/nrg3772.

- Pocha, C., Dieperink, E., McMaken, K. A., Knott, A., Thuras, P. and Ho, S. B. (2013) 'Surveillance for hepatocellular cancer with ultrasonography vs. computed tomography - A randomised study', *Alimentary Pharmacology and Therapeutics*. Aliment Pharmacol Ther, 38(3), pp. 303–312. doi: 10.1111/apt.12370.
- van der Pol, Y. and Mouliere, F. (2019) 'Toward the Early Detection of Cancer by Decoding the Epigenetic and Environmental Fingerprints of Cell-Free DNA.', *Cancer cell*, 36(4), pp. 350–368. doi: 10.1016/j.ccell.2019.09.003.
- Quinlan, A. R. and Hall, I. M. (2010) 'BEDTools: A flexible suite of utilities for comparing genomic features', *Bioinformatics*, 26(6), pp. 841–842. doi: 10.1093/bioinformatics/btq033.
- Raiber, E.-A., Hardisty, R., van Delft, P. and Balasubramanian, S. (2017) 'Mapping and elucidating the function of modified bases in DNA', *Nature Reviews Chemistry*. Springer Science and Business Media LLC, 1(9). doi: 10.1038/s41570-017-0069.
- Raisner, R., Kharbanda, S., Jin, L., Jeng, E., Chan, E., Merchant, M., Haverty, P. M., Bainer, R., Cheung, T., Arnott, D., Flynn, E. M., Romero, F. A., Magnuson, S. and Gascoigne, K. E. (2018) 'Enhancer Activity Requires CBP/P300 Bromodomain-Dependent Histone H3K27 Acetylation', *Cell Reports*. ElsevierCompany., 24(7), pp. 1722–1729. doi: 10.1016/j.celrep.2018.07.041.
- Ramakrishna, M., Williams, L. H., Boyle, S. E., Bearfoot, J. L., Sridhar, A., Speed, T. P., Gorringer, K. L. and Campbell, I. G. (2010) 'Identification of candidate growth promoting genes in ovarian cancer through integrated copy number and expression analysis', *PLoS ONE*, 5(4). doi: 10.1371/journal.pone.0009983.
- Ramírez, F., Dündar, F., Diehl, S., Grüning, B. A. and Manke, T. (2014) 'DeepTools: A

flexible platform for exploring deep-sequencing data', *Nucleic Acids Research*, 42(W1), pp. 187–191. doi: 10.1093/nar/gku365.

Rand, A. C., Jain, M., Eizenga, J. M., Musselman-Brown, A., Olsen, H. E., Akeson, M. and Paten, B. (2017) 'Mapping DNA methylation with high-throughput nanopore sequencing', *Nature Methods*. Nature Publishing Group, 14(4), pp. 411–413. doi: 10.1038/nmeth.4189.

Rauch, T. A., Zhong, X., Wu, X., Wang, M., Kernstine, K. H., Wang, Z., Riggs, A. D. and Pfeifer, G. P. (2008) 'High-resolution mapping of DNA hypermethylation and hypomethylation in lung cancer', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 105(1), pp. 252–257. doi: 10.1073/pnas.0710735105.

Rawla, P., Sunkara, T. and Gaduputi, V. (2019) 'Epidemiology of Pancreatic Cancer: Global Trends, Etiology and Risk Factors', *World Journal of Oncology*. Elmer Press, Inc., 10(1), pp. 10–27. doi: 10.14740/wjon1166.

Redrejo-Rodríguez, M., Vigouroux, A., Mursalimov, A., Grin, I., Alili, D., Koshenov, Z., Akishev, Z., Maksimenko, A., Bissenbaev, A. K., Matkarimov, B. T., Saparbaev, M., Ishchenko, A. A. and Moréra, S. (2016) 'Structural comparison of AP endonucleases from the exonuclease III family reveals new amino acid residues in human AP endonuclease 1 that are involved in incision of damaged DNA', *Biochimie*. Elsevier B.V., 128–129, pp. 20–33. doi: 10.1016/j.biochi.2016.06.011.

Riethdorf, S., Fritsche, H., Müller, V., Rau, T., Schindlbeck, C., Rack, B., Janni, W., Coith, C., Beck, K., Jänicke, F., Jackson, S., Gornet, T., Cristofanilli, M. and Pantel, K. (2007) 'Detection of circulating tumor cells in peripheral blood of patients with

metastatic breast cancer: A validation study of the cell search system', *Clinical Cancer Research*. Clin Cancer Res, 13(3), pp. 920–928. doi: 10.1158/1078-0432.CCR-06-1695.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W. and Smyth, G. K. (2015) 'Limma powers differential expression analyses for RNA-sequencing and microarray studies', *Nucleic Acids Research*, 43(7), p. e47. doi: 10.1093/nar/gkv007.

Rivera, C. M. and Ren, B. (2013) 'Mapping human epigenomes', *Cell*. Elsevier, pp. 39–55. doi: 10.1016/j.cell.2013.09.011.

Roberts, C., Raner, G. and Isaacs, G. (2018) 'High Performance Liquid Chromatography Separation of Epigenetic Cytosine Variants', *Methods and Protocols*. MDPI AG, 1(2), p. 10. doi: 10.3390/mps1020010.

Robinson, M. D., Stirzaker, C., Statham, A. L., Coolen, M. W., Song, J. Z., Nair, S. S., Strbenac, D., Speed, T. P. and Clark, S. J. (2010) 'Evaluation of affinity-based genome-wide DNA methylation data: Effects of CpG density, amplification bias, and copy number variation', *Genome Research*, 20(12), pp. 1719–1729. doi: 10.1101/gr.110601.110.

Rodriguez, J., Frigola, J., Vendrell, E., Risques, R. A., Fraga, M. F., Morales, C., Moreno, V., Esteller, M., Capellà, G., Ribas, M. and Peinado, M. A. (2006) 'Chromosomal instability correlates with genome-wide DNA demethylation in human primary colorectal cancers', *Cancer Research*. American Association for Cancer Research, 66(17), pp. 8462–8468. doi: 10.1158/0008-5472.CAN-06-0293.

Roller, A., Grossmann, V., Bacher, U., Poetzinger, F., Weissmann, S., Nadarajah, N., Boeck, L., Kern, W., Haferlach, C., Schnittger, S., Haferlach, T. and Kohlmann, A. (2013) 'Landmark analysis of DNMT3A mutations in hematological malignancies',

Leukemia. Nature Publishing Group, pp. 1573–1578. doi: 10.1038/leu.2013.65.

Rosenbloom, K. R., Armstrong, J., Barber, G. P., Casper, J., Clawson, H., Diekhans, M., Dreszer, T. R., Fujita, P. A., Guruvadoo, L., Haeussler, M., Harte, R. A., Heitner, S., Hickey, G., Hinrichs, A. S., Hubley, R., Karolchik, D., Learned, K., Lee, B. T., Li, C. H., Miga, K. H., Nguyen, N., Paten, B., Raney, B. J., Smit, A. F. A., Speir, M. L., Zweig, A. S., Haussler, D., Kuhn, R. M. and Kent, W. J. (2015) ‘The UCSC Genome Browser database: 2015 update’, *Nucleic Acids Research*. Oxford University Press, 43(D1), pp. D670–D681. doi: 10.1093/nar/gku1177.

Rothwell, D. G., Ayub, M., Cook, N., Thistlethwaite, F., Carter, L., Dean, E., Smith, N., Villa, S., Dransfield, J., Clipson, A., White, D., Nessa, K., Ferdous, S., Howell, M., Gupta, A., Kilerci, B., Mohan, S., Frese, K., Gulati, S., Miller, C., Jordan, A., Eaton, H., Hickson, N., O’Brien, C., Graham, D., Kelly, C., Aruketty, S., Metcalf, R., Chiramel, J., Tinsley, N., Vickers, A. J., Kurup, R., Frost, H., Stevenson, J., Southam, S., Landers, D., Wallace, A., Marais, R., Hughes, A. M., Brady, G., Dive, C. and Krebs, M. G. (2019) ‘Utility of ctDNA to support patient selection for early phase clinical trials: the TARGET study’, *Nature Medicine*. Nature Publishing Group, 25(5), pp. 738–743. doi: 10.1038/s41591-019-0380-z.

Saghafinia, S., Mina, M., Riggi, N., Hanahan, D. and Ciriello, G. (2018) ‘Pan-Cancer Landscape of Aberrant DNA Methylation across Human Tumors’, *Cell Reports*, October, pp. 1066-1080.e8. doi: 10.1016/j.celrep.2018.09.082.

Santos, F., Hendrich, B., Reik, W. and Dean, W. (2002) ‘Dynamic reprogramming of DNA methylation in the early mouse embryo’, *Developmental Biology*. Academic Press Inc., 241(1), pp. 172–182. doi: 10.1006/dbio.2001.0501.

Sato, N., Fukushima, N., Hruban, R. H. and Goggins, M. (2008) 'CpG island methylation profile of pancreatic intraepithelial neoplasia', *Modern Pathology*. NIH Public Access, 21(3), pp. 238–244. doi: 10.1038/modpathol.3800991.

Sato, N., Ueki, T., Fukushima, N., Iacobuzio–Donahue, C. A., Hruban, R. H., Goggins, M., Yeo, C. J. and Cameron, J. L. (2002) 'Aberrant methylation of CpG islands in intraductal papillary mucinous neoplasms of the pancreas', *Gastroenterology*. *Gastroenterology*, 123(1), pp. 365–372. doi: 10.1053/gast.2002.34160.

Schultz, M. D., He, Y., Whitaker, J. W., Hariharan, M., Mukamel, E. A., Leung, D., Rajagopal, N., Nery, J. R., Urich, M. A., Chen, H., Lin, S., Lin, Y., Jung, I., Schmitt, A. D., Selvaraj, S., Ren, B., Sejnowski, T. J., Wang, W. and Ecker, J. R. (2015) 'Human body epigenome maps reveal noncanonical DNA methylation variation', *Nature*, 523(7559), pp. 212–216. doi: 10.1038/nature14465.

Schulze, K., Nault, J. C. and Villanueva, A. (2016) 'Genetic profiling of hepatocellular carcinoma using next-generation sequencing', *Journal of Hepatology*. European Association for the Study of the Liver, 65(5), pp. 1031–1042. doi: 10.1016/j.jhep.2016.05.035.

Schutsky, E. K., DeNizio, J. E., Hu, P., Liu, M. Y., Nabel, C. S., Fabianic, E. B., Hwang, Y., Bushman, F. D., Wu, H. and Kohli, R. M. (2018) 'Nondestructive, base-resolution sequencing of 5-hydroxymethylcytosine using a DNA deaminase', *Nature biotechnology*, p. (in press). doi: 10.1038/nbt.4204.

Schwarze, K., Buchanan, J., Fermont, J. M., Dreau, H., Tilley, M. W., Taylor, J. M. J. C. J. M. J. C., Antoniou, P., Knight, S. J. L., Camps, C., Pentony, M. M., Kvikstad, E. M., Harris, S., Popitsch, N., Pagnamenta, A. T., Schuh, A., Taylor, J. M. J. C. J. M. J. C. and

Wordsworth, S. (2019) ‘The complete costs of genome sequencing: a microcosting study in cancer and rare diseases from a single center in the United Kingdom’, *Genetics in Medicine*. Springer US, 22(1), pp. 85–94. doi: 10.1038/s41436-019-0618-7.

Shah, C., Mramba, L. K., Bishnoi, R., Bejjanki, H., Chhatrala, H. S. and Chandana, S. R. (2017) ‘Survival differences among patients with hepatocellular carcinoma based on the stage of disease and therapy received: Pre and post sorafenib era’, *Journal of Gastrointestinal Oncology*. AME Publishing Company, 8(5), pp. 789–798. doi: 10.21037/jgo.2017.06.16.

Shapiro, R., DiFate, V. and Welcher, M. (1974) ‘Deamination Cytosine Derivatives by Bisulfite. Mechanism of the Reaction’, *Journal of the American Chemical Society*, 96(3), pp. 906–912. doi: 10.1021/ja00810a043.

Sharp, A. J., Stathaki, E., Migliavacca, E., Brahmachary, M., Montgomery, S. B., Dupre, Y. and Antonarakis, S. E. (2011) ‘DNA methylation profiles of human active and inactive X chromosomes’, *Genome Research*. Cold Spring Harbor Laboratory Press, 21(10), pp. 1592–1600. doi: 10.1101/gr.112680.110.

Shen, L., Wu, H., Diep, D., Yamaguchi, S., D’Alessio, A. C., Fung, H. L., Zhang, K. and Zhang, Y. (2013) ‘Genome-wide analysis reveals TET- and TDG-dependent 5-methylcytosine oxidation dynamics’, *Cell*, 153(3), pp. 692–706. doi: 10.1016/j.cell.2013.04.002.

Shen, S. Y., Singhanian, R., Fehring, G., Chakravarthy, A., Roehrl, M. H. A., Chadwick, D., Zuzarte, P. C., Borgida, A., Wang, T. T., Li, T., Kis, O., Zhao, Z., Spreafico, A., Medina, T. da S., Wang, Y., Roulois, D., Ettayebi, I., Chen, Z., Chow, S., Murphy, T., Arruda, A., O’Kane, G. M., Liu, J., Mansour, M., McPherson, J. D., O’Brien, C., Leighl,

N., Bedard, P. L., Fleshner, N., Liu, G., Minden, M. D., Gallinger, S., Goldenberg, A., Pugh, Trevor J., Hoffman, M. M., Bratman, S. V, Hung, rayjean J. and De Carvalho, D. D. (2018) 'Sensitive tumour detection and classification using plasma cell-free DNA methylomes', *Nature*. doi: 10.1038/s41586-018-0703-0.

Shinjo, K., Hara, K., Nagae, G., Umeda, T., Katsushima, K., Suzuki, M., Murofushi, Y., Umezui, Y., Takeuchi, I., Takahashi, S., Okuno, Y., Matsuo, K., Ito, H., Tajima, S., Aburatani, H., Yamao, K. and Kondo, Y. (2020) 'A novel sensitive detection method for DNA methylation in circulating free DNA of pancreatic cancer', *Plos One*, 15(6), p. e0233782. doi: 10.1371/journal.pone.0233782.

Shirley, M. (2020) 'Epi proColon® for Colorectal Cancer Screening: A Profile of Its Use in the USA', *Molecular Diagnosis & Therapy*. Mol Diagn Ther. doi: 10.1007/s40291-020-00473-8.

Siravegna, G., Marsoni, S., Siena, S. and Bardelli, A. (2017) 'Integrating liquid biopsies into the management of cancer', *Nature Reviews Clinical Oncology*, pp. 531–548. doi: 10.1038/nrclinonc.2017.14.

Skvortsova, K., Masle-Farquhar, E., Luu, P. L., Song, J. Z., Qu, W., Zotenko, E., Gould, C. M., Du, Q., Peters, T. J., Colino-Sanguino, Y., Pidsley, R., Nair, S. S., Khoury, A., Smith, G. C., Miosge, L. A., Reed, J. H., Kench, J. G., Rubin, M. A., Horvath, L., Bogdanovic, O., Lim, S. M., Polo, J. M., Goodnow, C. C., Stirzaker, C. and Clark, S. J. (2019) 'DNA Hypermethylation Encroachment at CpG Island Borders in Cancer Is Predisposed by H3K4 Monomethylation Patterns', *Cancer Cell*. Elsevier, 35(2), pp. 297-314.e8. doi: 10.1016/j.ccell.2019.01.004.

Smallwood, S. A., Lee, H. J., Angermueller, C., Krueger, F., Saadeh, H., Peat, J.,

Andrews, S. R., Stegle, O., Reik, W. and Kelsey, G. (2014) ‘Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity’, *Nature Methods*, 11(8), pp. 817–820. doi: 10.1038/nmeth.3035.

Smith, Z. D. and Meissner, A. (2013) ‘DNA methylation: Roles in mammalian development’, *Nature Reviews Genetics*, pp. 204–220. doi: 10.1038/nrg3354.

Song, C.-X., Yin, S., Ma, L., Wheeler, A., Chen, Y., Zhang, Y., Liu, B., Xiong, J., Zhang, W., Hu, J., Zhou, Z., Dong, B., Tian, Z., Jeffrey, S. S., Chua, M.-S., So, S., Li, W., Wei, Y., Diao, J., Xie, D. and Quake, S. R. (2017) ‘5-Hydroxymethylcytosine signatures in cell-free DNA provide information about tumor types and stages.’, *Cell research*. Nature Publishing Group, 27(10), pp. 1231–1242. doi: 10.1038/cr.2017.106.

Song, C. X., Szulwach, K. E., Fu, Y., Dai, Q., Yi, C., Li, X., Li, Y., Chen, C. H., Zhang, W., Jian, X., Wang, J., Zhang, L., Looney, T. J., Zhang, B., Godley, L. A., Hicks, L. M., Lahn, B. T., Jin, P. and He, C. (2011) ‘Selective chemical labeling reveals the genome-wide distribution of 5-hydroxymethylcytosine’, *Nature Biotechnology*, 29(1), pp. 68–75. doi: 10.1038/nbt.1732.

Song, L., James, S. R., Kazim, L. and Karpf, A. R. (2005) ‘Specific method for the determination of genomic DNA methylation by liquid chromatography-electrospray ionization tandem mass spectrometry’, *Analytical Chemistry*, 77(2), pp. 504–510. doi: 10.1021/ac0489420.

Sorenson, G. D., Pribish, D. M., Valone, F. H., Memoli, V. A., Bzik, D. J. and Yao, S. L. (1994) ‘Soluble Normal and Mutated DNA Sequences from Single-Copy Genes in Human Blood’, *Cancer Epidemiology Biomarkers and Prevention*, 3(1), pp. 67–71.

Spruijt, C. G., Gnerlich, F., Smits, A. H., Pfaffeneder, T., Jansen, P. W. T. C., Bauer, C.,

Münzel, M., Wagner, M., Müller, M., Khan, F., Eberl, H. C., Mensinga, A., Brinkman, A. B., Lephikova, K., Müller, U., Walter, J., Boelens, R., Van Ingen, H., Leonhardt, H., Carell, T. and Vermeulen, M. (2013) 'Dynamic readers for 5-(Hydroxy)methylcytosine and its oxidized derivatives', *Cell*. Elsevier, 152(5), pp. 1146–1159. doi: 10.1016/j.cell.2013.02.004.

Stadler, M. B., Murr, R., Burger, L., Ivanek, R., Lienert, F., Schöler, A., Wirbelauer, C., Oakeley, E. J., Gaidatzis, D., Tiwari, V. K. and Schübeler, D. (2011) 'DNA-binding factors shape the mouse methylome at distal regulatory regions', *Nature*. Nature Publishing Group, 480(7378), pp. 490–495. doi: 10.1038/nature10716.

Stark, A. P., Sacks, G. D., Rochefort, M. M., Donahue, T. R., Reber, H. A., Tomlinson, J. S., Dawson, D. W., Eibl, G. and Hines, O. J. (2016) 'Long-term survival in patients with pancreatic ductal adenocarcinoma', *Surgery*, 159(6), pp. 1520–1527. doi: 10.1016/j.surg.2015.12.024.

Sun, K., Jiang, P., Chan, K. C. A., Wong, J., Cheng, Y. K. Y., Liang, R. H. S., Chan, W. K., Ma, E. S. K., Chan, S. L., Cheng, S. H., Chan, R. W. Y., Tong, Y. K., Ng, S. S. M., Wong, R. S. M., Hui, D. S. C., Leung, T. N., Leung, T. Y., Lai, P. B. S., Chiu, R. W. K. and Lo, Y. M. D. (2015) 'Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 112(40), pp. E5503–E5512. doi: 10.1073/pnas.1508736112.

Sun, K., Jiang, P., Cheng, S. H., Cheng, T. H. T., Wong, J., Wong, V. W. S., Ng, S. S. M., Ma, B. B. Y., Leung, T. Y., Chan, S. L. Y., Mok, T. S. K., Lai, P. B. S., Chan, H. L. Y., Sun, H., Allen Chan, K. C., Chiu, R. W. K. and Dennis Lo, Y. M. (2019) 'Orientation-

aware plasma cell-free DNA fragmentation analysis in open chromatin regions informs tissue of origin', *Genome Research*. Cold Spring Harbor Laboratory Press, 29(3), pp. 418–427. doi: 10.1101/gr.242719.118.

Surveillance, Epidemiology, and End Results Program (2020). Available at: <https://seer.cancer.gov/> (Accessed: 2 July 2020).

Susan, J. Ci., Harrison, J., Paul, C. L. and Frommer, M. (1994) 'High sensitivity mapping of methylated cytosines', *Nucleic Acids Research*, 22(15), pp. 2990–2997. doi: 10.1093/nar/22.15.2990.

Szjli, E., Fehr, T. and Medzihradsky, K. F. (2008) 'Investigating the quantitative nature of MALDI-TOF MS', *Molecular and Cellular Proteomics*. American Society for Biochemistry and Molecular Biology, 7(12), pp. 2410–2418. doi: 10.1074/mcp.M800108-MCP200.

Tahiliani, M., Koh, K. P., Shen, Y., Pastor, W. A., Bandukwala, H., Brudno, Y., Agarwal, S., Iyer, L. M., Liu, D. R., Aravind, L. and Rao, A. (2009) 'Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1', *Science*. American Association for the Advancement of Science, 324(5929), pp. 930–935. doi: 10.1126/science.1170116.

Taiwo, O., Wilson, G. A., Morris, T., Seisenberger, S., Reik, W., Pearce, D., Beck, S. and Butcher, L. M. (2012) 'Methylome analysis using MeDIP-seq with low DNA concentrations', *Nature Protocols*. Nature Publishing Group, 7(4), pp. 617–636. doi: 10.1038/nprot.2012.012.

Tamanaha, E., Guan, S., Marks, K. and Saleh, L. (2016) 'Distributive Processing by the Iron(II)/ α -Ketoglutarate-Dependent Catalytic Domains of the TET Enzymes Is

Consistent with Epigenetic Roles for Oxidized 5-Methylcytosine Bases', *Journal of the American Chemical Society*. UTC, 138(30), pp. 9345–9348. doi: 10.1021/jacs.6b03243.

Tan, M., Luo, H., Lee, S., Jin, F., Yang, J. S., Montellier, E., Buchou, T., Cheng, Z., Rousseaux, S., Rajagopal, N., Lu, Z., Ye, Z., Zhu, Q., Wysocka, J., Ye, Y., Khochbin, S., Ren, B. and Zhao, Y. (2011) 'Identification of 67 histone marks and histone lysine crotonylation as a new type of histone modification', *Cell*. NIH Public Access, 146(6), pp. 1016–1028. doi: 10.1016/j.cell.2011.08.008.

Tanaka, K. and Okamoto, A. (2007) 'Degradation of DNA by bisulfite treatment', *Bioorganic and Medicinal Chemistry Letters*, 17(7), pp. 1912–1915. doi: 10.1016/j.bmcl.2007.01.040.

Tang, B., Li, Y., Qi, G., Yuan, S., Wang, Z., Yu, S., Li, B. and He, S. (2015) 'Clinicopathological Significance of CDKN2A Promoter Hypermethylation Frequency with Pancreatic Cancer', *Scientific Reports*. Nature Publishing Group, 5(1), p. 13563. doi: 10.1038/srep13563.

Terragni, J., Bitinaite, J., Zheng, Y. and Pradhan, S. (2012) 'Biochemical characterization of recombinant β -glucosyltransferase and analysis of global 5-Hydroxymethylcytosine in unique genomes', *Biochemistry*, 51(5), pp. 1009–1019. doi: 10.1021/bi2014739.

Teschendorff, A. E., Jones, A., Fiegl, H., Sargent, A., Zhuang, J. J., Kitchener, H. C. and Widschwendter, M. (2012) 'Epigenetic variability in cells of normal cytology is associated with the risk of future morphological transformation', *Genome Medicine*. BioMed Central, 4(3), p. 24. doi: 10.1186/gm323.

Teschendorff, A. E., Menon, U., Gentry-Maharaj, A., Ramus, S. J., Weisenberger, D. J., Shen, H., Campan, M., Noushmehr, H., Bell, C. G., Maxwell, A. P., Savage, D. A.,

Mueller-Holzner, E., Marth, C., Kocjan, G., Gayther, S. A., Jones, A., Beck, S., Wagner, W., Laird, P. W., Jacobs, I. J. and Widschwendter, M. (2010) 'Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer', *Genome Research*. Cold Spring Harbor Laboratory Press, 20(4), pp. 440–446. doi: 10.1101/gr.103606.109.

Thierry, A. R., El Messaoudi, S., Gahan, P. B., Anker, P. and Stroun, M. (2016) 'Origins, structures, and functions of circulating DNA in oncology', *Cancer and Metastasis Reviews*, 35(3), pp. 347–376. doi: 10.1007/s10555-016-9629-x.

Thomson, J. P., Hunter, J. M., Nestor, C. E., Dunican, D. S., Terranova, R., Moggs, J. G. and Meehan, R. R. (2013) 'Comparative analysis of affinity-based 5-hydroxymethylation enrichment techniques', *Nucleic Acids Research*, 41(22), pp. 1–15. doi: 10.1093/nar/gkt1080.

Tian, Z., Tolić, N., Zhao, R., Moore, R. J., Hengel, S. M., Robinson, E. W., Stenoien, D. L., Wu, S., Smith, R. D. and Paša-Tolić, L. (2012) 'Enhanced top-down characterization of histone post-translational modifications', *Genome Biology*. BioMed Central, 13(10), p. R86. doi: 10.1186/gb-2012-13-10-R86.

Torres, S., González, Á., Cunqueiro Tomas, A. J., Calabuig Fariñas, S., Ferrero, M., Mirda, D., Sirera, R., Jantus-Lewintre, E. and Camps, C. (2020) 'A profile on cobas® EGFR Mutation Test v2 as companion diagnostic for first-line treatment of patients with non-small cell lung cancer', *Expert Review of Molecular Diagnostics*. Taylor and Francis Ltd, 20(6). doi: 10.1080/14737159.2020.1724094.

Trevisani, F., D'Intino, P. E., Morselli-Labate, A. M., Mazzella, G., Accogli, E., Caraceni, P., Domenicali, M., De Notariis, S., Roda, E. and Bernardi, M. (2001) 'Serum

α -fetoprotein for diagnosis of hepatocellular carcinoma in patients with chronic liver disease: Influence of HBsAg and anti-HCV status', *Journal of Hepatology*, 34(4), pp. 570–575. doi: 10.1016/S0168-8278(00)00053-2.

Tucci, V., Isles, A. R., Kelsey, G., Ferguson-Smith, A. C., Bartolomei, M. S., Benvenisty, N., Bourc'his, D., Charalambous, M., Dulac, C., Feil, R., Glaser, J., Huelsmann, L., John, R. M., McNamara, G. I., Moorwood, K., Muscatelli, F., Sasaki, H., Strassmann, B. I., Vincenz, C. and Wilkins, J. (2019) 'Genomic Imprinting and Physiological Processes in Mammals', *Cell*, pp. 952–965. doi: 10.1016/j.cell.2019.01.043.

Turcan, S., Rohle, D., Goenka, A., Walsh, L. A., Fang, F., Yilmaz, E., Campos, C., Fabius, A. W. M., Lu, C., Ward, P. S., Thompson, C. B., Kaufman, A., Guryanova, O., Levine, R., Heguy, A., Viale, A., Morris, L. G. T., Huse, J. T., Mellinghoff, I. K. and Chan, T. A. (2012) 'IDH1 mutation is sufficient to establish the glioma hypermethylator phenotype', *Nature*. NIH Public Access, 483(7390), pp. 479–483. doi: 10.1038/nature10866.

Turner, N. C., Kingston, B., Kilburn, L. S., Kernaghan, S., Wardley, A. M., Macpherson, I. R., Baird, R. D., Roylance, R., Stephens, P., Oikonomidou, O., Braybrooke, J. P., Tuthill, M., Abraham, J., Winter, M. C., Bye, H., Hubank, M., Gevensleben, H., Cutts, R., Snowdon, C., Rea, D., Cameron, D., Shaaban, A., Randle, K., Martin, S., Wilkinson, K., Moretti, L., Bliss, J. M. and Ring, A. (2020) 'Circulating tumour DNA analysis to direct therapy in advanced breast cancer (plasmaMATCH): a multicentre, multicohort, phase 2a, platform trial', *The Lancet Oncology*. Elsevier, 0(0). doi: 10.1016/S1470-2045(20)30444-7.

Tzatsos, A., Paskaleva, P., Ferrari, F., Deshpande, V., Stoykova, S., Contino, G., Wong, K. K., Lan, F., Trojer, P., Park, P. J. and Bardeesy, N. (2013) 'KDM2B promotes

pancreatic cancer via Polycomb-dependent and -independent transcriptional programs’, *Journal of Clinical Investigation*. J Clin Invest, 123(2), pp. 727–739. doi: 10.1172/JCI64535.

Underhill, H. R., Kitzman, J. O., Hellwig, S., Welker, N. C., Daza, R., Baker, D. N., Gligorich, K. M., Rostomily, R. C., Bronner, M. P. and Shendure, J. (2016) ‘Fragment Length of Circulating Tumor DNA’, *PLoS Genetics*, 12(7), p. 1006162. doi: 10.1371/journal.pgen.1006162.

Vacca, M., Della Ragione, F., Scalabrì, F. and D’Esposito, M. (2016) ‘X inactivation and reactivation in X-linked diseases’, *Seminars in Cell and Developmental Biology*. Academic Press, pp. 78–87. doi: 10.1016/j.semcdb.2016.03.009.

Vaisvila, R., Ponnaluri, V. K. C., Sun, Z., Langhorst, B. W., Saleh, L., Guan, S., Dai, N., Campbell, M. A., Sexton, B., Marks, K., Samaranayake, M., Samuelson, J. C., Church, H. E., Tamanaha, E., Corrêa, I. R., Pradhan, S., Dimalanta, E. T., Evans, T. C., Williams, L. and Davis, T. B. (2019) ‘EM-seq: Detection of DNA Methylation at Single Base Resolution from Picograms of DNA’, *bioRxiv*. Cold Spring Harbor Laboratory, p. 2019.12.20.884692. doi: 10.1101/2019.12.20.884692.

Vasioukhin, V., Anker, P., Maurice, P., Lyautey, J., Lederrey, C. and Stroun, M. (1994) ‘Point mutations of the N-ras gene in the blood plasma DNA of patients with myelodysplastic syndrome or acute myelogenous leukaemia’, *British Journal of Haematology*. John Wiley & Sons, Ltd, 86(4), pp. 774–779. doi: <https://doi.org/10.1111/j.1365-2141.1994.tb04828.x>.

Venkhataraman, R., Donald, C. D., Roy, R., You, H. J., Doetsch, P. W. and Kow, Y. W. (2001) ‘Enzymatic processing of DNA containing tandem dihydrouracil by

endonucleases III and VIII', *Nucleic Acids Res.*, 29(2), pp. 407–414. Available at: <http://nar.oupjournals.org/cgi/content/abstract/29/2/407>.

Vidal, E., Sayols, S., Moran, S., Guillaumet-Adkins, A., Schroeder, M. P., Royo, R., Orozco, M., Gut, M., Gut, I., Lopez-Bigas, N., Heyn, H. and Esteller, M. (2017) 'A DNA methylation map of human cancer at single base-pair resolution', *Oncogene*, 36(40), pp. 5648–5657. doi: 10.1038/onc.2017.176.

Vincent, A., Omura, N., Hong, S. M., Jaffe, A., Eshleman, J. and Goggins, M. (2011) 'Genome-wide analysis of promoter methylation associated with gene expression profile in pancreatic adenocarcinoma', *Clinical Cancer Research*. American Association for Cancer Research, 17(13), pp. 4341–4354. doi: 10.1158/1078-0432.CCR-10-3431.

Wan, J. C. M., Heider, K., Gale, D., Murphy, S., Fisher, E., Mouliere, F., Ruiz-Valdepenas, A., Santonja, A., Morris, J., Chandrananda, D., Marshall, A., Gill, A. B., Chan, P. Y., Barker, E., Young, G., Cooper, W. N., Hudecova, I., Marass, F., Mair, R., Brindle, K. M., Stewart, G. D., Abraham, J. E., Caldas, C., Rassl, D. M., Rintoul, R. C., Alifrangis, C., Middleton, M. R., Gallagher, F. A., Parkinson, C., Durrani, A., McDermott, U., Smith, C. G., Massie, C., Corrie, P. G. and Rosenfeld, N. (2020) 'ctDNA monitoring using patient-specific sequencing and integration of variant reads', *Science Translational Medicine*. American Association for the Advancement of Science, 12(548). doi: 10.1126/scitranslmed.aaz8084.

Wan, J. C. M., Massie, C., Garcia-Corbacho, J., Mouliere, F., Brenton, J. D., Caldas, C., Pacey, S., Baird, R. and Rosenfeld, N. (2017) 'Liquid biopsies come of age: Towards implementation of circulating tumour DNA', *Nature Reviews Cancer*. Nature Publishing Group, pp. 223–238. doi: 10.1038/nrc.2017.7.

Wang, J., Xia, Y., Li, L., Gong, D., Yao, Y., Luo, H., Lu, H., Yi, N., Wu, H., Zhang, X., Tao, Q. and Gao, F. (2013) 'Double restriction-enzyme digestion improves the coverage and accuracy of genome-wide CpG methylation profiling by reduced representation bisulfite sequencing', *BMC Genomics*. BioMed Central, 14(1), p. 11. doi: 10.1186/1471-2164-14-11.

Wang, L., Zhou, Y., Xu, L., Xiao, R., Lu, X., Chen, L., Chong, J., Li, H., He, C., Fu, X. D. and Wang, D. (2015) 'Molecular basis for 5-carboxycytosine recognition by RNA polymerase II elongation complex', *Nature*. Nature Publishing Group, 523(7562), pp. 621–625. doi: 10.1038/nature14482.

Wang, Y., Guo, H., Zhang, D., Yu, X., Leng, X., Li, S. and Zhu, W. (2016) 'Overexpression of SOX18 correlates with accelerated cell growth and poor prognosis in human pancreatic ductal adenocarcinoma', *Biochemical and Biophysical Research Communications*. Elsevier B.V., 479(3), pp. 510–516. doi: 10.1016/j.bbrc.2016.09.099.

Wang, Z., Yin, J., Zhou, W., Bai, J., Xie, Y., Xu, K., Zheng, X., Xiao, J., Zhou, L., Qi, X., Li, Y., Li, X. and Xu, J. (2020) 'Complex impact of DNA methylation on transcriptional dysregulation across 22 human cancer types', *Nucleic Acids Research*, 48(5), pp. 2287–2302. doi: 10.1093/nar/gkaa041.

Watson, J. D. and Crick, F. H. C. (1953) 'Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid', *Nature*, 171(4356), pp. 737–738. doi: 10.1038/171737a0.

Watts, G. (2018) 'Liquid biopsy: still early days for early detection', *Lancet (London, England)*, pp. 2593–2594. doi: 10.1016/S0140-6736(18)31476-4.

Wenger, A. M., Peluso, P., Rowell, W. J., Chang, P. C., Hall, R. J., Concepcion, G. T., Ebler, J., Fungtammasan, A., Kolesnikov, A., Olson, N. D., Töpfer, A., Alonge, M.,

Mahmoud, M., Qian, Y., Chin, C. S., Phillippy, A. M., Schatz, M. C., Myers, G., DePristo, M. A., Ruan, J., Marschall, T., Sedlazeck, F. J., Zook, J. M., Li, H., Koren, S., Carroll, A., Rank, D. R. and Hunkapiller, M. W. (2019) ‘Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome’, *Nature Biotechnology*. Nature Publishing Group, 37(10), pp. 1155–1162. doi: 10.1038/s41587-019-0217-9.

Wetterstrand KA (2016) *DNA Sequencing Costs: Data, National Human Genome Research Institute*. Available at: <https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data> (Accessed: 13 September 2020).

Whittaker, S., Marais, R. and Zhu, A. X. (2010) ‘The role of signaling pathways in the development and treatment of hepatocellular carcinoma’, *Oncogene*. Nature Publishing Group, 29(36), pp. 4989–5005. doi: 10.1038/onc.2010.236.

Wilhelm-Benartzi, C. S., Koestler, D. C., Karagas, M. R., Flanagan, J. M., Christensen, B. C., Kelsey, K. T., Marsit, C. J., Houseman, E. A. and Brown, R. (2013) ‘Review of processing and analysis methods for DNA methylation array data’, *British Journal of Cancer*. Nature Publishing Group, pp. 1394–1402. doi: 10.1038/bjc.2013.496.

Witte, T., Plass, C. and Gerhauser, C. (2014) ‘Pan-cancer patterns of DNA methylation’, *Genome Medicine*. BioMed Central, p. 66. doi: 10.1186/s13073-014-0066-6.

Worm Ørntoft, M. B., Jensen, S. Ø., Hansen, T. B., Bramsen, J. B. and Andersen, C. L. (2017) ‘Comparative analysis of 12 different kits for bisulfite conversion of circulating cell-free DNA’, *Epigenetics*. Taylor and Francis Inc., 12(8), pp. 626–636. doi: 10.1080/15592294.2017.1334024.

Wu, H., Wu, X. and Zhang, Y. (2016) ‘Base-resolution profiling of active DNA

demethylation using MAB-seq and caMAB-seq', *Nature Protocols*, 11(6), pp. 1081–1100. doi: 10.1038/nprot.2016.069.

Wu, H. and Zhang, Y. (2014) 'Reversing DNA methylation: Mechanisms, genomics, and biological functions', *Cell*. Elsevier, pp. 45–68. doi: 10.1016/j.cell.2013.12.019.

Wutz, A. and Valencia, K. (2015) 'Recent insights into the regulation of X-chromosome inactivation', *Advances in Genomics and Genetics*. Dove Medical Press Ltd., 5, p. 227. doi: 10.2147/agg.s60399.

Xia, B., Han, D., Lu, X., Sun, Z., Zhou, A., Yin, Q., Zeng, H., Liu, M., Jiang, X., Xie, W., He, C. and Yi, C. (2015) 'Bisulfite-free, base-resolution analysis of 5-formylcytosine at the genome scale', *Nature Methods*. Nature Publishing Group, pp. 1047–1050. doi: 10.1038/nmeth.3569.

Xiang, S. and Tong, L. (2008) 'Crystal structures of human and *Staphylococcus aureus* pyruvate carboxylase and molecular insights into the carboxyltransfer reaction', *Nature Structural and Molecular Biology*. Nat Struct Mol Biol, 15(3), pp. 295–302. doi: 10.1038/nsmb.1393.

Xiao, F. H., Wang, H. T. and Kong, Q. P. (2019) 'Dynamic DNA methylation during aging: A “prophet” of age-related outcomes', *Frontiers in Genetics*. Frontiers Media S.A. doi: 10.3389/fgene.2019.00107.

Xie, W., Barr, C. L., Kim, A., Yue, F., Lee, A. Y., Eubanks, J., Dempster, E. L. and Ren, B. (2012) 'Base-resolution analyses of sequence and parent-of-origin dependent DNA methylation in the mouse genome', *Cell*. Elsevier, 148(4), pp. 816–831. doi: 10.1016/j.cell.2011.12.035.

Xie, W., Kagiampakis, I., Pan, L., Zhang, Y. W., Murphy, L., Tao, Y., Kong, X., Kang, B., Xia, L., Carvalho, F. L. F., Sen, S., Chiu Yen, R.-W., Zahnow, C. A., Ahuja, N., Baylin, S. B. and Easwaran, H. (2018) 'DNA Methylation Patterns Separate Senescence from Transformation Potential and Indicate Cancer Risk.', *Cancer cell*. Elsevier, 33(2), pp. 309-321.e5. doi: 10.1016/j.ccell.2018.01.008.

Xu, Y., Li, X., Yang, Y., Li, C. and Shao, X. (2019) 'Human age prediction based on DNA methylation of non-blood tissues', *Computer Methods and Programs in Biomedicine*. Elsevier Ireland Ltd, 171, pp. 11–18. doi: 10.1016/j.cmpb.2019.02.010.

Yang, B., Guo, M., Herman, J. G. and Clark, D. P. (2003) 'Aberrant promoter methylation profiles of tumor suppressor genes in hepatocellular carcinoma', *American Journal of Pathology*. American Society for Investigative Pathology, 163(3), pp. 1101–1107. doi: 10.1016/S0002-9440(10)63469-4.

Yang, L., Chen, Z., Stout, E. S., Delerue, F., Ittner, L. M., Wilkins, M. R., Quinlan, K. G. R. and Crossley, M. (2020) 'Methylation of a CGATA element inhibits binding and regulation by GATA-1.', *Nature communications*. Nature Publishing Group, 11(1), p. 2560. doi: 10.1038/s41467-020-16388-1.

Yao, M. and Kow, Y. W. (1997) 'Further characterization of Escherichia coli endonuclease V. Mechanism of recognition for deoxyinosine, deoxyuridine, and base mismatches in DNA', *Journal of Biological Chemistry*. American Society for Biochemistry and Molecular Biology, 272(49), pp. 30774–30779. doi: 10.1074/jbc.272.49.30774.

Yasuda, A., Sawai, H., Takahashi, H., Ochi, N., Matsuo, Y., Funahashi, H., Sato, M., Okada, Y., Takeyama, H. and Manabe, T. (2006) 'The stem cell factor/c-kit receptor

pathway enhances proliferation and invasion of pancreatic cancer cells’, *Molecular Cancer*. BioMed Central, 5, p. 46. doi: 10.1186/1476-4598-5-46.

Yegnasubramanian, S., Haffner, M. C., Zhang, Y., Gurel, B., Cornish, T. C., Wu, Z., Irizarry, R. A., Morgan, J., Hicks, J., Deweese, T. L., Isaacs, W. B., Steven Bova, G., De Marzo, A. M. and Nelson, W. G. (2008) ‘DNA Hypomethylation Arises Later in Prostate Cancer Progression than CpG Island Hypermethylation and Contributes to Metastatic Tumor Heterogeneity’, *Cancer Res*, 68(21), pp. 8954–67. doi: 10.1158/0008-5472.CAN-07-6088.

Yin, Y., Morgunova, E., Jolma, A., Kaasinen, E., Sahu, B., Khund-Sayeed, S., Das, P. K., Kivioja, T., Dave, K., Zhong, F., Nitta, K. R., Taipale, M., Popov, A., Ginno, P. A., Domcke, S., Yan, J., Schübeler, D., Vinson, C. and Taipale, J. (2017) ‘Impact of cytosine methylation on DNA binding specificities of human transcription factors’, *Science*. American Association for the Advancement of Science, 356(6337). doi: 10.1126/science.aaj2239.

Yu, J.-S., Chen, Y.-T., Chiang, W.-F., Hsiao, Y.-C., Chu, L. J., See, L.-C., Wu, C.-S., Tu, H.-T., Chen, H.-W., Chen, C.-C., Liao, W.-C., Chang, Y.-T., Wu, C.-C., Lin, C.-Y., Liu, S.-Y., Chiou, S.-T., Chia, S.-L., Chang, K.-P., Chien, C.-Y., Chang, S.-W., Chang, C.-J., Young, J. D., Pao, C. C., Chang, Y.-S. and Hartwell, L. H. (2016) ‘Saliva protein biomarkers to detect oral squamous cell carcinoma in a high-risk population in Taiwan’, *Proceedings of the National Academy of Sciences*, 113(41), pp. 11549–11554. doi: 10.1073/pnas.1612368113.

Yu, M., Hon, G. C., Szulwach, K. E., Song, C. X., Zhang, L., Kim, A., Li, X., Dai, Q., Shen, Y., Park, B., Min, J. H., Jin, P., Ren, B. and He, C. (2012) ‘Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome’, *Cell*, 149(6), pp. 1368–1380.

doi: 10.1016/j.cell.2012.04.027.

Zeng, H., He, B., Xia, B., Dongsheng, B., Lu, X., Cai, J., Chen, L., Zhou, A., Zhu, C., Meng, H., Gao, Y., Guo, H., He, C., Dai, Q. and Yi, C. (2018) 'Bisulfite-free, nano-scale analysis of 5-hydroxymethylcytosine at single base resolution', *Journal of the American Chemical Society*, p. jacs.8b08297. doi: 10.1021/jacs.8b08297.

Zeng, H., He, B. and Yi, C. (2019) 'Compilation of Modern Technologies To Map Genome-Wide Cytosine Modifications in DNA', *ChemBioChem*, 20(15), pp. 1898–1905. doi: 10.1002/cbic.201900035.

Zhang, C., Li, Z., Cheng, Y., Jia, F., Li, R., Wu, M., Li, K. and Wei, L. (2007) 'CpG island methylator phenotype association with elevated serum α -fetoprotein level in hepatocellular carcinoma', *Clinical Cancer Research*. Clin Cancer Res, 13(3), pp. 944–952. doi: 10.1158/1078-0432.CCR-06-2268.

Zhang, G. and Pradhan, S. (2014) 'Mammalian epigenetic mechanisms', *IUBMB Life*, 66(4), pp. 240–256. doi: 10.1002/iub.1264.

Zhang, L., Chen, W., Iyer, L. M., Hu, J., Wang, G., Fu, Y., Yu, M., Dai, Q., Aravind, L. and He, C. (2014) 'A TET homologue protein from *Coprinopsis cinerea* (CcTET) that biochemically converts 5-methylcytosine to 5-hydroxymethylcytosine, 5-formylcytosine, and 5-carboxylcytosine', *Journal of the American Chemical Society*. American Chemical Society, 136(13), pp. 4801–4804. doi: 10.1021/ja500979k.

Zhang, L., Szulwach, K. E., Hon, G. C., Song, C. X., Park, B., Yu, M., Lu, X., Dai, Q., Wang, X., Street, C. R., Tan, H., Min, J. H., Ren, B., Jin, P. and He, C. (2013) 'Tet-mediated covalent labelling of 5-methylcytosine for its genome-wide detection and sequencing', *Nature Communications*, 4. doi: 10.1038/ncomms2527.

Zhang, M., Wu, J., Zhong, W., Zhao, Z. and Liu, Z. (2018) ‘Long non-coding RNA BRE-AS1 represses non-small cell lung cancer cell growth and survival via up-regulating NR4A3’, *Archives of Biochemistry and Biophysics*. Academic Press Inc., 660, pp. 53–63. doi: 10.1016/j.abb.2018.09.013.

Zhang, Y., Petropoulos, S., Liu, J., Cheishvili, D., Zhou, R., Dymov, S., Li, K., Li, N. and Szyf, M. (2018) ‘The signature of liver cancer in immune cells DNA methylation’, *Clinical Epigenetics*. BioMed Central Ltd., 10(1), p. 8. doi: 10.1186/s13148-017-0436-1.

Zhou, W., Dinh, H. Q., Ramjan, Z., Weisenberger, D. J., Nicolet, C. M., Shen, H., Laird, P. W. and Berman, B. P. (2018) ‘DNA methylation loss in late-replicating domains is linked to mitotic cell division’, *Nature Genetics*. doi: 10.1038/s41588-018-0073-4.

Zhu, C., Gao, Y., Guo, H., Xia, B., Song, J., Wu, X., Zeng, H., Kee, K., Tang, F. and Yi, C. (2017) ‘Single-Cell 5-Formylcytosine Landscapes of Mammalian Early Embryos and ESCs at Single-Base Resolution’, *Cell Stem Cell*. Cell Press, 20(5), pp. 720-731.e5. doi: 10.1016/j.stem.2017.02.013.

Ziller, M. J., Gu, H., Müller, F., Donaghey, J., Tsai, L. T. Y., Kohlbacher, O., De Jager, P. L., Rosen, E. D., Bennett, D. A., Bernstein, B. E., Gnirke, A. and Meissner, A. (2013) ‘Charting a dynamic DNA methylation landscape of the human genome’, *Nature*, 500(7463), pp. 477–481. doi: 10.1038/nature12433.

Zviran, A., Schulman, Rafael C., Shah, M., Hill, S. T. K., Deochand, S., Khamnei, C. C., Maloney, D., Patel, K., Liao, W., Widman, A. J., Wong, P., Callahan, M. K., Ha, G., Reed, S., Rotem, D., Frederick, D., Sharova, T., Miao, B., Kim, T., Gydush, G., Rhoades, J., Huang, K. Y., Omans, N. D., Bolan, P. O., Lipsky, A. H., Ang, C., Malbari, M.,

Spinelli, C. F., Kazancioglu, S., Runnels, A. M., Fennessey, S., Stolte, C., Gaiti, F., Inghirami, G. G., Adalsteinsson, V., Houck-Loomis, B., Ishii, J., Wolchok, J. D., Boland, G., Robine, N., Altorki, N. K. and Landau, D. A. (2020) 'Genome-wide cell-free DNA mutational integration enables ultra-sensitive cancer monitoring', *Nature Medicine*. Springer US, 26(7), pp. 1114–1124. doi: 10.1038/s41591-020-0915-3.

Zviran, A., Schulman, Rafael C, Shah, M., Hill, S. T. K., Deochand, S., Khamnei, C. C., Maloney, D., Patel, K., Liao, W., Widman, A. J., Wong, P., Callahan, M. K., Ha, G., Reed, S., Rotem, D., Frederick, D., Sharova, T., Miao, B., Kim, T., Gydush, G., Rhoades, J., Huang, K. Y., Omans, N. D., Bolan, P. O., Lipsky, A. H., Ang, C., Malbari, M., Spinelli, C. F., Kazancioglu, S., Runnels, A. M., Fennessey, S., Stolte, C., Gaiti, F., Inghirami, G. G., Adalsteinsson, V., Houck-Loomis, B., Ishii, J., Wolchok, J. D., Boland, G., Robine, N., Altorki, N. K. and Landau, D. A. (2020) 'Genome-wide cell-free DNA mutational integration enables ultra-sensitive cancer monitoring', *Nature Medicine*, 26(7), pp. 1114–1124. doi: 10.1038/s41591-020-0915-3.

10 Supplementary Materials

10.1 Appendix 1 - Glossary

2-HG	2-hydroxyglutarate
5caC:	5-carboxylcytosine
5fC:	5-formylcytosine
5gmC:	glucosyl-5-hydroxymethyl-cytosine
5hmC:	5-hydroxymethylcytosine
5mC:	5-methylcytosine
ACE-seq:	APOBEC-coupled epigenetic sequencing
AFP:	alpha fetoprotein
APE1:	Apurinic/apyrimidinic Endonuclease 1
APOBEC:	apolipoprotein B mRNA editing catalytic polypeptide-like
AUC:	area under a receiver operating characteristic
AUROC:	area under a receiver operating characteristic
BER:	base excision repair
BS-seq:	Bisulfite Sequencing
CA19-9:	Carbohydrate antigen 19-9
CASP8:	Caspase 8
CCGA:	Circulating Cell-Free DNA Genome Atlas

ccTET:	<i>Coprinopsis Cinerea</i> TET
CDKN2A:	Cyclin Dependent Kinase Inhibitor 2A
cfDNA:	cell-free DNA
cfMeDIP:	cell-free methylated DNA immunoprecipitation and high-throughput sequencing
cfRNA:	cell-free RNA
CGI:	CpG island
CIMP:	CpG island methylator phenotype
CNV:	copy number variation
CRC:	colorectal cancer
CT:	Computed Tomography
CTC:	circulating tumour cell
ctDNA:	circulating tumour DNA
DBCO-PEG4-biotin:	dibenzocyclooctyne-PEG4-biotin
DELFI:	DNA evaluation of fragments for early interception
DHU:	dihydrouracil
DMC:	differentially methylated CpG site
DMR:	differentially methylated region
DNA:	deoxyribonucleic acid
DNMT:	DNA methyltransferase

DSBH:	double-stranded β -helix domain
EEIA:	enhanced enzyme-based immunoassay
eeTAPS:	Endonuclease enrichment TAPS
EGFR:	Epidermal Growth Factor Receptor
ELISA:	enzyme-linked immunosorbent assay
FAS:	Fas Cell Surface Death Receptor
FDA:	The Food and Drug Administration
FIT:	faecal immunochemical test
Fpg:	Formamidopyrimidine DNA Glycosylase
FPKM:	Fragments Per Kilobase of transcript per Million mapped reads
HBV:	hepatitis B virus
HCC:	hepatocellular carcinoma
HCV:	hepatitis C virus
HK2:	Hexokinase 2
HPLC-UV:	High Performance Liquid Chromatography-Ultraviolet
IDH1:	isocitrate dehydrogenase 1
IGF1:	Insulin Like Growth Factor 1
JBP2:	J binding protein 1
JBP2:	J binding protein 2
KDM2B:	Lysine Demethylase 2B

KDM5A:	Lysine Demethylase 5A
LC-MS/MS:	Liquid chromatography coupled with tandem mass spectrometry
LINE-1:	long interspersed element-1
lncRNA:	long non-coding RNA
LOD:	limit of detection
MALDI:	matrix-assisted laser desorption/ionization mass spectroscopy
MBD:	methyl-CpG binding domain
mCpG:	methyated CpG site
MeCP1:	Methyl-CpG-binding domain protein complex 1
MeDIP-seq:	Methylated DNA immunoprecipitation sequencing
MRD:	minimal residual disease
MRE-seq:	Methylation-Sensitive Restriction Enzyme Sequencing
MRI:	Magnetic Resonance Imaging
mTAG-seq:	Methyltransferase-directed Transfer of Activated Groups Sequencing
mTET1CD:	mouse ten-eleven translocation methylcytosine dioxygenases 1 catalytic domain
N ₃ -5gmC:	6-N ₃ -β-glucosyl-5-hydroxymethyl-cytosine
ncRNA:	non-coding RNA
NgTET1	<i>Naegleria guberi</i> TET1
NSCLC:	non-small-cell lung carcinoma

oxBS-seq:	Oxidative Bisulfite Sequencing
PanIN:	pancreatic intraepithelial neoplasia
PC:	Pyruvate Carboxylase
PCA:	Principal component analysis
PCR:	Polymerase chain reaction
PDAC:	pancreatic ductal adenocarcinoma
PET:	Positron emission tomography
pic-BH ₃ :	2-picoline borane (pic-borane)
PyBr:	pyridine borane
qPCR:	Quantitative Polymerase chain reaction
RCC:	renal cell carcinoma
RNA:	Ribonucleic acid
RRBS:	Reduced Representation Bisulfite Sequencing
rrTAPS:	Reduced Representation TAPS
SAH:	S-Adenosyl-L-homocysteine
SAM:	S-Adenosyl-L-methionine
SINE:	Short interspersed nuclear elements
SMRT-seq:	Single-molecule, real-time sequencing
STR:	short tandem repeat
T4-BGT:	T4-betaglucosyltransferase

TAB-seq:	Tet-assisted bisulfite sequencing
TDG:	Thymine-DNA glycosylase
TES:	Transcription end site
TET:	ten-eleven translocation methylcytosine dioxygenases
TF:	transcription factor
TGF- β :	tumour growth factor beta
TKI:	tyrosine kinase inhibitor
TLC:	layer chromatography
TSS:	Transcription start site
UDG:	Uracil-DNA glycosylase
UDP:	uridine diphosphate
UDP-6-N ₃ Glc:	UDP-6-azide-glucose
umCpG:	unmethylated CpG site
USER:	Uracil-specific excision reagent
WGBS:	whole genome bisulfite sequencing
wgTAPS:	whole genome TAPS
XCI:	X chromosome inactivation
α KG:	alpha ketoglutarate

10.2 Appendix 2 - Publications

Y. Liu*, **P. Siejka-Zielińska***, G. Velikova, Y. Bi, F. Yuan, M. Tomkova, C. Bai, L. Chen, B. Schuster-Böckler, C.-X. Song. Bisulfite-free direct detection of 5-methylcytosine and 5-hydroxymethylcytosine at base resolution. *Nat. Biotechnol.* (2019), doi:10.1038/s41587-019-0041-2.

F. Yuan, Y. Bi, **P. Siejka-Zielinska**, Y.-L. Zhou, X.-X. Zhang, C.-X. Song. Bisulfite-free and base-resolution analysis of 5-methylcytidine and 5-hydroxymethylcytidine in RNA with peroxotungstate. *Chem. Commun.* **55** (2019), doi:10.1039/c9cc00274j.

Y. Liu, J. Cheng, **P. Siejka-Zielińska**, C. Weldon, H. Roberts, M. Lopopolo, A. Magri, V. D'Arienzo, J. M. Harris, J. A. McKeating, C. X. Song. Accurate targeted long-read DNA methylation and hydroxymethylation sequencing with TAPS. *Genome Biol.* **21**, 54 (2020).

J. Cheng*, **P. Siejka-Zielińska***, Y. Liu, A. Chandran, S. Kriaucionis, C.X. Song. Endonuclease enrichment TAPS for cost-effective genome-wide base-resolution DNA methylation detection. *In preparation*.

* These authors contributed equally

10.3 Appendix 3 – Supplementary Materials to Chapter 3



Supplementary Figure 3-1. TOF-MS analysis of NgTET1 A212V protein. Molecular weight of major peak is consistent with expected molecular weight of NgTET1 A212V. Wild-type NgTET1 or other protein contaminants was not detected in analysed protein fraction.

Supplementary Table 3-1. Sequencing and mapping statistics of cfMeSeal libraries.

Sample ID	Total number of reads	Uniquely mapped, deduplicated reads	Uniquely mapped, deduplicated reads (%)
colon16	28243964	14502300	51
colon17	35430593	18363990	52
colon19	12302458	8510810	69
GBM57	17684038	12347131	70
GBM66	26905906	19686506	73
pancreatic9	15412626	10872529	71
pancreatic15	16722243	11244532	67
pancreatic22	20325591	12660586	62
pancreatic27	19000238	13257428	70
pancreatic68	23432557	15407976	66
pancreatic69	17367658	12303401	71
pancreatic75	18146716	12543402	69
stomach2	16690918	11279941	68
stomach3	17851706	12542714	70
stomach8	24021397	13166902	55
H9	19892970	14191164	71
H10_rep1	29463200	17078038	58
H10_rep2	28693930	16910164	59
H1083	28338690	15718368	55
H2652	29525962	16989417	58
H2654	24491701	15237535	62
H2657	35459330	18427340	52
H3455	34481216	17937204	52

Supplementary Table 3-2. Details of spike-in controls for cfMeSeal method. Spike-in controls were amplified from lambda DNA. HMC and MC control were amplified in presence of 5% 5hmC or 5mC, respectively. Remaining controls were amplified using only unmodified cytosines. CpG_1, CpG_2, CpG_8, CpG_16, CpG_19 controls were CpG-methylated by M.SssI methyltransferase.

Spike-in ID	Sequence (5' → 3')	Primers (5' → 3')
C	CGTTTCCGTTCTTCTTCGTCATAACTT AATGTTTTTATTTAAAATACCCTCTGA AAAGAAAGGAAACGACAGGTGCTGA AAGCGAGGCTTTTTGGCCTCTGTGCTT TCCTTTCTCTGTTTTTGTCCGTGGAAT GAACAATGGAAGTCAACAAAAAGCA GCTGGCTGACATTTTCGGTGCGAGTA	C_Fwd CGTTTCCGTTCTTCTTCGTC C_Rev TACTCGCACCGAAAATGTCA
HMC	TGAAAACGAAAGGGGATACGGGAAA ACGTAAACCTTCGTAAACACCACGC TCGGTGAGACGTGGGAGGCGAAAATT GGCGAACGTCCGGATGCTGAAGTGAT GGCAGAGCGGAAAGAGCATTATTCA GCGCCCGTTCCTGACCGTGTGGCTTA CCTGACCGCCGGTATCGACTCCCAGC TGGAC	HMC_Fwd TGAAAACGAAAGGGGATACG HMC_Rev GTCCAGCTGGGAGTCGATAC
MC	GTGGCGGGTTATGATGAACTTGCTGC TTTTGATGATGATATTGAACAGGAAG GCTCTCCGACGTTCTGGGTGACAAG CGTATTGAAGGCTCGGTCTGGCCAAA GTCCATCCGTGGCTCCACGCCAAAAG TGAGAGGCACCTGTCAGATTGAGCGT GCAGCCAGTGAATCCCCGCATTTTAT G	MC_Fwd GTGGCGGGTTATGATGAACT MC_Rev CATAAAATGCGGGGATTAC
CpG_1	TTGGCCATACTACTAAATCCTGAATA GCTTTAAGAAGGTTATGTTTAAACC ATCGCTTAATTTGCTGAGATTAACAT AGTAGTCAATGCTTTCACCTAAGGAA AAAAACATTTAGGGAGTTGACTGAA TTTTTTATCTATTAATGAATAAGTGCT TACTTCTTCTTTTGGACC	CpG_1_Fwd TTGGCCATACTACTAAATCCTG CpG_1_Rev GGTCAAAAAGAAGAAGTAAGCAC
CpG_2	AGCTTCAAGCCAGAGTTGTCTTTTTCT ATCTACTCTCATACAACCAATAAATG CTGAAATGAATTCTAAGCGGAGATCG CCTAGTGATTTTAACTATTGCTGGC AGCATTCTTGAGTCCAATATAAAAGT ATTGTGTACCTTTTGCTGGGTCAGGTT GTTCT	CpG_2_Fwd AGCTTCAAGCCAGAGTTGTC CpG_2_Rev AGAACAACCTGACCCAGC
CpG_5	CCTGATGAAACAAGCATGTCATCGTA ATATGTTCTAGCGGGTTTGTTTTATC TCGGAGATTATTTTCATAAAGCTTTTC TAATTTAACCTTTGTCAGGTTACCAAC TACTAAGGTTGTAGGCTCAAGAGGGT GTGTCCTGTCGTAGGTAAATAACTGA CCTGTCGAGCTTAATATTCTATATTGT TGTTCTTTCTGCAAAAAAGTGGGGAA GTGAGTAATG	CpG_5_Fwd CCTGATGAAACAAGCATGTC CpG_5_Rev CATTACTCACTTCCCCACTT
CpG_8	CAATGCCACAAAGAAGAGTCAATCGC AGACAACATTTTGAATGCGGTCACAC	CpG_8_Fwd CAATGCCACAAAGAAGAGTC

	GTTAGCAGCATGATTGCCACGGATGG CAACATATTAACGGCATGATATTGAC TTATTGAATAAAATTGGGTAAATTTG ACTCAACGATGGGTAAATTCGCTCGT TGTGGTAGTGAGATGAAAAGAGG	CpG_8_Rev CCTCTTTTCATCTCACTACC
CpG_16	TGGCTGGGTCAGCAGTATCGGTAAGG CGGTGACGGCGAAGGAAGTGATCAC CCGCACGGTGAAAGTCACCAATGTGG GACGTCCGTCGATGGCAGAAGATCGC AGCACGGTAACAGCGGCAACCGGCA TGACCGTGACGCCTGCCAGCACCTCG GTGGTGAAAGGGCAGAGCACCA	CpG_16_Fwd TGGCTGGGTCAGCAGTATCG CpG_16_Rev TGGTGCTCTGCCCTTTCAC
CpG_19	GGAAGCTGACCACAGACCCGACGTCG GGGCTGACGGCGATGGCTCGCCAGTT CCATAACGTGTCGGCGGAGCAGATTG CGTATGTTGCTCAGTTGCAGCGTTCC GGCGATGAAGCCGGGGCATTGCAGG CGGCGAACGAGGCCGCAACGAAAGG GTTTGATGACCAGACC	CpG_19_Fwd GGAAGCTGACCACAGACC PS_CG19_Rev GGTCTGGTCATCAAACCCT

Supplementary Table 3-3. Recombinant NgTET1 details. NgTET1 insert was cloned between BamHI and XhoI restriction sites of plasmid pRSET-A.

Protein	Protein Length	Protein MW	Tag	Plasmid	Insert length
NgTET1	321 aa	36498.34 Da	His-tag	pRSET-A	966
ccDNA Sequence (w/o tag)					
ATGACCACCTTTAAACAGCAGACCATCAAAGAGAAAGAAACCAAACGCAAATAC TGCATTAAAGGCACCACCGCAAATCTGACCCAGACACATCCGAATGGTCCGTTT GTGTTAATCGTGGTGAAGAAGTTGCAAATACCACCACACTGCTGGATAGCGGTG GTGGCATTAAACAAAAAAGCCTGCTGCAGAATCTGCTGAGCAAATGTAAACCA CCTTTCAGCAGAGCTTTACCAATGCAAATATCACCTGAAAGATGAGAAATGGCT GAAAAATGTTTCGCACCGCCTATTTTGTGTTGTGATCATGATGGTAGCGTTGAACTG GCATATCTGCCGAATGTTCTGCCGAAAGAACTGGTTGAAGAATTCACCGAAAAA TTCGAAAGCATTTCAGACCGGTCGTAAAAAAGATACCGGTTATAGCGGTATTCTGG ATAATTCAATGCCGTTCAATTATGTGACCGCAGATCTGAGCCAAGAAGCTGGGTCA GTATCTGAGCGAAATTGTTAATCCGCAGATCAACTACTACATCAGCAAAGCTGCTG ACCTGTGTTAGCAGCCGTACCATTAACTATCTGGTTAGCCTGAACGATAGCTATT ATGCCCTGAATAATTGTCTGTATCCGAGCACCGCATTTAATAGCCTGAAACCGAG CAATGATGGTCATCGTATTCGTAAACCGCATAAAGATAACCTGGATATTACCCCG AGCAGCCTGTTTTATTTTCGGCAATTTTCAGAACACCGAAGGCTATCTGGAAGTGA CCGATAAAAACTGCAAAGTTTTTGTTCAGCCTGGTGATGTGCTGTTTTTCAAAGG CAACGAATATAAACATGTGGTGGCCAATATTACCAGCGGTTGGCGTATTGGTCTG GTTTATTTTGCACATAAAGGCAGCAAAACCAAACCGTATTATGAGGATACCCAG AAGAACAGCCTGAAGATTCATAAAGAAACGAAATAA					
Protein Sequence (w/o tag)					
MTTFKQQTIKEKETKRKYCIKGTTANLTQTHPNPVCVNRGEEVANTTTLLDSGGI NKKSLQNLKSKCTTFQQSFTNANITLKDEKWLKNVRTAYFVCDHDGSVELAYLP NVLPKELVEEFTEKFESIQTGRKKDTGYSGILDNSMPFNYVTADLSQELGQYLSEIVN PQINYYISKLLTCVSSRTINYLVSLNDSYYALNNCLYPSTAFNSLKPSNDGHRIRKPHK DNLDTTPSSLFYFGNFQNTGYLELTDKNCKVFVQPGDVLFFKGNEYKHVVANITSG WRIGLVYFAHKGSKTKPYEDTQKNSLKIHKETK					
Tag sequence			Protein length (with tag)	Protein MW (with tag)	
MRGSHHHHHHGMASMTGGQMQGRDLYDDDDKDRWGS			358 aa	40627.77 Da	

Supplementary Table 3-4. Recombinant ccTET details. ccTET insert was cloned between BamHI and XhoI restriction sites of plasmid pRSET-A.

Protein	Protein Length	Protein MW	Tag	Plasmid	Insert length
ccTET	430 aa	48058.18 Da	His-tag	pRSET-A	1293
ccDNA Sequence (w/o tag)					
ATGAGCGCAATTCCGTTTAGCACCGATGATTGTAGCCAGGATGAAACCCTGCCGA GCCTGCTGCTGATTGATGAAGCAGCAGCAGTTCTGGGTCGTATGATTCAGGGTCT GCGTACCGGTATTCCGTATATTCATACCGAAAACGATAGCATTAAAGCGAATCCG ATTCTGCGCACCGCACTGTGGCAGGCAGCATATGTTCTGGAAAAAGCATATCGTC GTCGTTATCGTGTTCCGTGGACCGCACGTCGTTATATGCGTGAAGTACACCGCG TCAGGATGGTCGTAATGCAAATCGTGAAGCAGTTATGGCAAAGAATTTCCGCCT GGTGCAGAACTGAATAGCGATCATCCGGTTCAAGAAATTCTGCCAGCAATGATT ATTGATGCCGAAGATCATATCCTGTTTTGTTATCTGCCGAGCTGTGTTAGTCCGGC AATTATGACCATTATCGATGCAGCAGTTGGCACCCCTGGCAACCACCAAAGATGGT CATCTGCAGAAAAAAGCCGTGCACGTGAAGGTGAACGTGCCCCGTGTTGAAATG CAGAAAGTTAAAGGTAAAGGCAAACAAGAAGAAGAGGGCCAAGAAAAACTGGG TGCAAATTGGCGTGAAGCACTGGACCTGTTTCGTCAGGGTGCATGTAAATGACA CCGGGTGTTCTGACCTTTGCACCGGCATGGTGGCCTGTTGGTCATGAAAATCAGC TGCCTGGTCCGGCAAGCACCCCTGAAACCGCCTAAAGGTGAAGGTCGTATGTTTCT GAGCGATATTCCGATTGCAAGCGCACTGGTTGGTGCAATTCTGGCACAGATTAAT CAGCCGCTGTTTGAAAGCGGTGTTAAAGTTCTGCGCGAACTGTATAGCAATAGCA AACTGACCAAAGATCATAGCACCGTGAGCAAAATTATCGAAATTTGGTTTAGCCC GTTTAGCTCACTGAGCCTGATTGTTAATCGTGCAACCCCGATTATCGTGATACC AGCGGTCCGATTGAAGGTATGGATATTCTGGTTACCGGTGGCAATTATTCAAATG GTGTTCTGGTGACCCCGAGCTTTAATCGTCGTTGGACCTATAATCCGGGTTGTGTT GTTGCACTGCTGGGTAAACTGGTTCTGCATGGTGTTCGGGAAGTTGATGGTGAAC GTTATTGTATGGCACATTTTTGGCGTGAGCGTCTGTTTGATGCAGCCGGTGTTCCG TTCCGTATCCGAGCAAATGGCAAGAAAGCTATACCAATATTCCGGAAGCGCTGC AGTGGCACCGTGATAATGTTACCCGTTAA					
Protein Sequence (w/o tag)					
MSAIPFSTDDCSQDETLPSLLIDEAAAVLGRMIQGLRTGIPYIHTENDSIKANPILRTA LWQAAAYVLEKAYRRRYRVPWTARRYMRELTPRQDGRNANREAVMAKEFPFGAEL NSDHPVQEILPAMIIDAEDHILFCYLPSCVSPAIMTIIDAAVGTLATTKDGHLLQKKSRA REGERARVEMQKVKGKQKEEGQEKLGANWREALDLFRQGACKMTPGVLTFAF AWWPVGHENQLPGPASTLKPCKGEGRMFLSDIPIASALVGAILAQINQLFESGVKVL RELYSNSKLTKDHSTVSKIIEIWFSPFSSLSLIVNRATPIHRDTSGPIEGMDILVTGGNYS NGVLVTPSFNRRWTYNPGCVVALLGKLVLHGVPEVDGERYCMAMHFWRERLFDAAG VPFPYPSKWQESYTNPEALQWHRDNVTR					
Tag sequence			Protein length (with tag)	Protein MW (with tag)	
MRGSHHHHHHGMASMTGGQMQGRDLYDDDDKDRWGS			466 aa	52187.53 Da	

Supplementary Table 3-5. Recombinant NgTET1 A212V details. Mutation site is bolded.

Protein	Protein Length	Protein MW	Tag	Plasmid	Insert length
NgTET1 A212V	321 aa	36526.39 Da	His-tag	pRSET-A	966
ccDNA Sequence (w/o tag)					
ATGACCACCTTTAAACAGCAGACCATCAAAGAGAAAGAAACCAAACGCAAATAC TGCATTAAAGGCACCACCGCAAATCTGACCCAGACACATCCGAATGGTCCGGTTT GTGTTAATCGTGGTGAAGAAGTTGCAAATACCACCACACTGCTGGATAGCGGTG GTGGCATTAAACAAAAAAGCCTGCTGCAGAATCTGCTGAGCAAATGTAAACCA CCTTTCAGCAGAGCTTTACCAATGCAAATATCACCTGAAAGATGAGAAATGGCT GAAAAATGTTTCGCACCGCCTATTTTGTGTTGTGATCATGATGGTAGCGTTGAACTG GCATATCTGCCGAATGTTCTGCCGAAAGAACTGGTTGAAGAATTCACCGAAAAA TTCGAAAGCATTCAGACCGGTCGTAAAAAAGATACCGGTTATAGCGGTATTCTGG ATAATTCAATGCCGTTCAATTATGTGACCGCAGATCTGAGCCAAGAAGTGGGTCA GTATCTGAGCGAAATTGTTAATCCGCAGATCAACTACTACATCAGCAAAGTGTG ACCTGTGTTAGCAGCCGTACCATTAAGTATCTGGTTAGCCTGAACGATAGCTATT ATGCCCTGAATAATTGTCTGTATCCGAGCACCG GTG TTTAATAGCCTGAAACCGAG CAATGATGGTCATCGTATTCGTAAACCGCATAAAGATAACCTGGATATTACCCCG AGCAGCCTGTTTTATTTTCGGCAATTTTCAGAACACCGAAGGCTATCTGGAAGTGA CCGATAAAAACTGCAAAGTTTTTGTTCAGCCTGGTGATGTGCTGTTTTTCAAAGG CAACGAATATAACATGTGGTGGCCAATATTACCAGCGGTTGGCGTATTGGTCTG GTTTATTTTGCACATAAAGGCAGCAAAACCAAACCGTATTATGAGGATACCCAG AAGAACAGCCTGAAGATTCATAAAGAAACGAAATAA					
Protein Sequence (w/o tag)					
MTTFKQQTIKEKETKRKYCIKGTTANLTQTHPNGPVCVNRGEEVANTTTLLDSGGI NKKSLLQNLLSKCKTTFQQSFTNANITLKDEKWLKNVRTAYFVCDHDGSVELAYLP NVLPKELVEEFTEKFESIQTGRKKDTGYSGILDNSMPFNYVTADLSQELGQYLSEIVN PQINYYISKLLTCVSSRTINYLVSLNDSYYALNNCLYPSTVFNSLKPSNDGHRIRKPHK DNLDTSSSLFYFGNFQNTGYLELTDKNCKVFVQPGDVLFFKGNEYKHVVANITSG WRIGLVYFAHKGSKTKPYEDTQKNSLKIHKETK					
Tag sequence				Protein length (with tag)	Protein MW (with tag)
MRGSHHHHHHGMASMTGGQQMGRDLYDDDDKDRWGS				358 aa	40655.82 Da

Supplementary Table 3-6. Detailed sequence of 100 bp model DNA.

Strand	Sequence (5' → 3')
Top strand	ACTCTCACTCTCACCTCCATCTTACTTTCTACTATCTACG TA5hm CTCAT TTCGATATCAATTGTATACTACCTATCTACATACACGAAGTGGTAGGTA AGGG
Bottom strand	CCCTTACCTACCACTTCGTGTATGTAGATAGGTAGTATACAATTGATAT CGAAATGAGTACGTAGATAGTAGAAAGTAAGATGGAGGTGAGAGTGA GAGT

Supplementary Table 3-7. Sequences of qPCR primers for 5mC pull-down validation.

Spike-in ID	Forward primer (5' → 3')	Reverse primer (5' → 3')
C	GCCTCTGTCGTTTCCTTTCT	ATGTCAGCCAGCTGCTTT
HMC	ATGCTGAAGTGATGGCAGAG	ATACCGGCGGTCAGGTAA
MC	CCTGGGTGACAAGCGTATT	GACAGGTGCCTCTCACTTT
CpG_1	AAGGTTATGTTTAAAACCATCGCT	CCTTAGGTGAAAGCATTGACTACT
CpG_2	CTAAGCGGAGATCGCCTAGTG	TGACCCAGCAAAAGGTACAC
CpG_5	GCGGGTTTGTTTTTATCTCGG	AGTTGGTAACCTGACAAAGGT
CpG_8	TGCCACGGATGGCAACATA	CGAGCGAATTAACCCATCGT
CpG_16	AAGGAAGTGATCACCCGCAC	TACCGTGCTGCGATCTTCTG
CpG_19	TATGTTGCTCAGTTGCAGCG	CAAACCCTTTCGTTGCGGC

10.4 Appendix 4 – Supplementary Materials to Chapter 4

Supplementary Table 4-1. Detailed sequences of 2kb spike-in control.

2kb
CACAGATGTCTGCCTGTTTCATCCGCGTCCAGCTCGTTGAGTTTCTCCAGAAGCGT TAATGTCTGGCTTCTGATAAAGCGGGCCATGTTAAGGGCGGTTTTTTCCTGTTTG GTCAGTATGCTCTCCGTGTAAGGGGGATTTCTGTTTCATGGGGGTAATGATACCGA TGAAACGAGAGAGGATGCTCACGATACGGGTACTGATGATGAACATGCCCGGT TACTGGAACGTTGTGAGGGTAAACAACCTGGCGGTATGGATGCGGCGGGACCAGA GAAAAATCACTCAGGGTCAATGCCAGCGCTTCGTTAATACAGATGTAGGTGTTCC ACAGGGTAGCCAGCAGCATCCTGCGATGCAGATCCGGAACATAATGGTGCAGGG CGCTGACTTCCGCGTTTTCCAGACTTTACGAAACACGGAACCGAAGACCATTCAT GTTGTTGCTCAGGTTCGACAGCTTTTGCAGCAGCAGTCGCTTCACGTTTCGCTCGC GTATCGGTGATTCATTCTGCTAACCCAGTAAGGCAACCCCGCCAGCCTAGCCGGGT CCTCAACGACAGGAGCACGATCATGCGCACCCGTGGGGCCGCCATGCCGGCGAT AATGGCCTGCTTCTCGCCGAAACGTTTGGTGGCGGGACCAGTGACGAAGGCTTG AGCGAGGGCGTGCAAGATTCCGAATACCGCAAGCGACAGGCCGATCATCGTCGC GCTCCAGCGAAAGCGGTCCTCGCCGAAATGACCCAGAGCGCTGCCGGCACCTG TCCTACGAGTTGCATGATAAAGAAGACAGTCATAAGTGCGGCGACGATAGTCAT GCCCCGCGCCACCGGAAGGAGCTGACTGGGTGAAGGCTCTCAAGGGCATCGG TCGAGATCCCGGTGCCTAATGAGTGAGCTAACTTACATTAATTGCGTTGCGCTCA CTGCCCGCTTTCCAGTCGGGAAACCTGTCGTGCCAGCTGCATTAATGAATCGGCC AACGCGCGGGGAGAGGCGGTTTTGCGTATTGGGCGCCAGGGTGGTTTTTCTTTTCA CCAGTGAGACGGGCAACAGCTGATTGCCCTTACCCGCTGGCCCTGAGAGAGTT GCAGCAAGCGGTCCACGCTGGTTTGCCCCAGCAGGCGAAAATCCTGTTTGATGG TGGTTAACGGCGGGATATAACATGAGCTGTCTTCGGTATCGTCGTATCCCACTAC CGAGATATCCGCACCAACGCGCAGCCCGGACTCGGTAATGGCGCGCATTGCGCC CAGCGCCATCTGATCGTTGGCAACCAGCATCGCAGTGGAACGATGCCCTCATTC AGCATTTGCATGGTTTGTGAAAACCGGACATGGCACTCCAGTCGCCTTCCCGTT CCGCTATCGGCTGAATTTGATTGCGAGTGAGATATTTATGCCAGCCAGCCAGACG CAGACGCGCCGAGACAGAACTTAATGGGCCCCGCTAACAGCGCGATTTGCTGGTG ACCCAATGCGACCAGATGCTCCACGCCCAGTCGCGTACCGTCTTCATGGGAGAA AATAATACTGTTGATGGGTGTCTGGTCAGAGACATCAAGAAATAACGCCGGAAC ATTAGTGACGGCAGCTTCCACAGCAATGGCATCCTGGTCATCCAGCGGATAGTTA ATGATCAGCCCACTGACGCGTTGCGCGAGAAGATTGTGCACCGCCGCTTTACAG GCTTCGACGCGGCTTCGTTCTACCATCGACACCACCACGCTGGCACCCAGTTGAT CGGCGCGAGATTTAATCGCCGCGACAATTTGCGACGGCGCGTGCAGGGCCAGAC TGGAGGTGGCAACGCCAATCAGCAACGACTGTTTGCCCGCCAGTTGTTGTGCCAC GCGGTTGGGAATGTAATTCAGCTCCGCCATCGCCGCTTCCACTTTTTCCCGCGTTT TCGCAGAAACGTGGCTGGCCTGGTTACCCACGCGGGAAACGGTCTGATAAGAGA CACCGGCATACTCTGCGACATCGTATAACGTTACTGGTTTCACATTCACCACCCT

Supplementary Table 4-2. Compound-dependent HPLC-MS/MS parameters used for nucleosides quantification. RT: retention time, CE: collision energy; CAE: cell accelerator voltage. All the nucleosides were analysed in the positive mode.

Compound	Precursor Ion (m/z)	Product Ion (m/z)	RT (min)	Delta RT (min)	CE (V)	CAE (V)
dA+H	252	136	13.78	2	10	4
dT+H	243	127	11.07	2	10	4
dT+Na	265	149	11.07	2	10	4
dG+H	268	152	9.64	2	10	4
dC+H	228	112	3.71	1.5	10	4
dC+Na	250	134	3.71	1.5	10	4
mdC+H	242	126	9.05	1.5	10	4
mdC+Na	264	148	9.05	1.5	10	4
hmdC+H	258	142	4.34	2	12	4
hmdC+Na	280	164	4.34	2	12	4
fdC+H	256	140	10.69	2	8	4
fdC+Na	278	162	10.69	2	8	4
cadC+H	272	156	1.75	3	12	4
cadC+Na	294	178	1.75	3	12	4
DHU+H	231	115	3.45	3	10	4

Supplementary Table 4-3. Recombinant mTet1CD details. mTet1CD insert was cloned between KpnI and BamHI restriction sites of plasmid pcDNA3-Flag.

Protein	Protein Length	Protein MW	Tag	Plasmid	Insert length
mTET1CD	682 aa	75398.02 Da	Flag-tag	pcDNA-Flag	2049
ccDNA Sequence (with tag)					
ATGGACTACAAAGACGATGACGACAAGGAAGCTGCACCCTGTGACTGTGATGGAGGTA CACAAAAAGAAAAAGGCCCATATTATACACACCTTGGGGCAGGACCAAGTGTGGCTGCT GTCAGGGAGCTCATGGAGACTAGGTTTGGCCAGAAGGGGAAGGCAATCCGGATTGAGAA GATAGTGTTACGGGGAAGGAAGGGAAGAGCTCTCAGGGCTGCCCGGTCGCCAAGTGGG TGATCAGAAGAAGTGGTCCTGAAGAGAAGCTTATTTGTTTGGTTCGTGAGCGTGTAGACC ATCACTGTTTCGACGGCTGTGATAGTTGTCCTTATCCTGCTGTGGGAAGGTATCCCTCGCCT GATGGCTGACCGCCTGTACAAAGAGCTCACTGAGAACTTGAGGTCCTACAGCGGACATCC CACAGACCGAAGATGTACCCTCAACAAAAAGCGTACCTGCACCTGTCAAGGCATCGACCC AAAAACCTGCGGAGCGTCCTTCTCCTTTGGCTGTTTCGTGGAGCATGTATTTCAACGGCTGT AAGTTTGGGAGGAGTGAAAACCCAGAAAATTACAGACTTGCTCCAAACTACCCCTTACAT AACTACTATAAGAGAATTACTGGAATGAGTTCTGAAGGAAGTGACGTGAAAACCGGGTG GATCATTCAGACCGCAAGACCCTCATAAGCAGAGAGGAAAAACAGCTTGAAAAGAATT TACAAGAATTGGCTACAGTATTAGCTCCACTTTACAAGCAGATGGCTCCAGTTGCTTATCA AAATCAGGTGGAATATGAAGAAGTTGCTGGAGACTGTCGACTTGGAATGAAGAGGGGC GTCCTTTCTCTGGTGTACCTGTTGCATGGATTTTTGTGCCATTCTCACAAGGACATTAC AACATGCACAACGGAAGCACCGTGGTGTGTACGTTGATTCGAGCAGATGGCCGTGACACA AATTGTCCCAGGATGAACAACCTCCACGTCCTGCCACTATACCGGCTTGACAGACTGAT GAATTTGGCTCCGTGGAAGGGATGAAGGCCAAAATCAAATCTGGGGCCATCCAAGTCAAT GGGCCAACAGGAAGAGGGCGACTACGTTTTACTGAGCCTGTTCTCGATGTGGGAAGAGG GCCAAAATGAAGCAGAACCACAATAAATCAGGTTACACAACTAAGAGCTTTTCATCA GCCTCATCTACTTCTCACCTAGTGAAAGACGAATCTACAGACTTCTGTCCCTGCAGGCTT CCTCCGCAGAAACATCTACCTGTACGTACAGTAAAACAGCCTCAGGTGGGTTTGCAGAAA CAAGTAGTATTCTCCACTGCACAATGCCTTCTGGAGCACACAGTGGTGCTAATGCAGCTG CTGGGGAATGTACTGGAACGGTGCAGCCTGCCGAGGTGGCTGCTCATCCTCACCAGTCTC TTCCACAGCCGATTCTCCCGTTCATGCTGAGCCTCTCACTAGTCCATCTGAGCAGCTAAC TTCTAACAGTCAAACCAGCAGCTCCCTCTCCTCAGCAATTCTCAGAAACTGGCTTCCTGT CAGGTGGAAGATGAGCGGCACCCTGAAGCGGATGAGCCTCAGCACCCCGAGGACGATAA CTTGCTCAACTTGATGAATTCTGGTCAGACAGTGAGGAGATCTACGCCGATCCTTCCTTT GGTGGCGTGGCGATAGCACCCATTACGGCTCGGTGCTCATTGAGTGCGCTCGGAAGGAG CTTCATGCTACCACCTCTTTGCGCTCCCCCAAACGAGGGGTCCCTTTTCGTGTGCCCTTGT ATTCTACCAGCACAAAAGCCTAAACAAGCCTAATCATGGTTTTGATATCAACAAAATTAA GTGTAAATGCAAAAAAGTAACGAAAAAAAAGCCCGCAGACCGGGAGTGTCTGATGTAT CCCCGAAGCCAATTTATCACACCAAAATTCCTTCTCGAGTTGCATCAACCTTAACCCGAGA CAATGTTGTTACCGTGTCCCCATACTCTCTCACTCATGTTGCGGGACCCTACAATCGTTGG GTCTAA					
Protein Sequence (with tag)					
MDYKDDDDKEAAPCDCDGGTQKEKGPYYTHLGAGPSVAAVRELMETRFGQKGKAIRIEKIV FTGKEGKSSQGCVPVAKWVIRRSGPPEEKLICLVREVDHHCSTAVIVVLILLWEGIPRLMADRL YKELTENLRSYSGHPTDRRCTLNKKRTCTCQGIDPKTCGASFSFGCSWSMYFNGCKFGRSENP RKFR LAPNYPLHNYK RITGMSSEGS DVKTGW IIPDRKTLISREEKQLEKNLQELATV LAPLYK QMAPVAYQNQVEYEEVAGDCRLGNEEGRPFSGVTCCMDFCASHKDIHNMHNGSTVVCTLI RADGRDTNCPEDQLHVLPLYRLADTDEFGSVEGMKAKIKSGAIQVNGPTRKRRLRFTEPVPR CGKRAKMKQNHKSGSHNTKSFSASSTSHLVKDESTDFCPLQASSAETSTCTYSKTASGGFA ETSSILHCTMPSGAHSGANAAAGECTGTVQPAEVAAPHQSLPTADSPVHAEPLTSPSEQLTS NQSNQQLPLLSNSQKLASCQVEDERHPEADEPQHPEDDNLPQLDEFWSDSEEIYADPSFGGVA IAPIHGSLVIECARKELHATTSLRSPKRGVPFRVSLVFYQHKSLNKP NHGFDINKIKCKCKKVT KKKPADRECPDVSPEANLSHQIPSRVASTLTRDNVTVSPYSLTHVAGPYNRWV					

Supplementary Table 4-4. Control sequences for TaqoI assay.

222mer
CCTGATGAAACAAGCATGTCATCGTAATATGTTCTAGCGGGTTTGTTTTATCTC GGAGATTATTTTCATAAAGCTTTTCTAATTTAACCTTTGTCAGGTTACCAACTACT AAGGTTGTAGGCTCAAGAGGGTGTGTCCTGTCGTAGGTAAATAACTGACCTGTC GAGCTTAATATTCTATATTGTTGTTCTTTCTGCAAAAAAGTGGGGAAGTGAGTAA TG
194mer – High GC control
GCTGGGGAACTACAGGCTGACAGTCCGGGCGGTAAATGCGTGGGGGCAGCAGG GCGATCCGGCGTCGGTATCGTTCCGGATTGCCGCACCGGCAGCACCGT TCGAGGA TTGAGCTGACGCCGGGCTATTTTCAGATAACCGCCACGCCGCATCTTGCCGTTTA TGACCCGACGGTACAGTTTGAGTTCTGGTTCT
216mer – Low GC control
GTCAGGTTACCAACTACTAAGGTTGTAGGCTCAAGAGGGTGTGTCCTGTCGTAGG TAAATAACTGACCTG TCGAGCTTAATATTCTATATTGTTGTTCTTTCTGCAAAAA AGTGGGGAAGTGAGTAATGAAATTATTTCTAACATTTATCTGCATCATACCTTCC GAGCATTTATTAAGCATTTTCGCTATAAGTTCTCGCTGGAAGAGGTAGTTTT

Supplementary Table 4-5. Mapping and sequencing quality statistics for TAPS and WGBS. Mapped reads are reads mapped to mm9+spike-ins+PhiX; MAPQ>0) Unique mapping rate was calculated as unique reads (MAPQ>0 / trimmed reads). Unique PCR deduplicated mapping rate was calculated as unique MAPQ>0 PCR deduplicated reads / trimmed reads).

Measure	TAPS	WGBS
Total raw reads	455548210	376062375
Trimmed reads	453028186	367860813
Mapped reads (for TAPS MAPQ>0)	400479295	254123777
PCR deduplicated reads	378138439	233568669
Unique mapping rate	88.99 %	69.08 %
Unique PCR deduplicated mapping rate	83.83 %	63.49 %

10.5 Appendix 5 – Supplementary Materials to Chapter 5

Supplementary Table 5-1. Mapping statistics of the pilot cfDNA TAPS libraries.

Sample	Total_trimmed	Mapped	% mapped	Deduplicated	% Deduplicated
Smp_1	160056546	147440929	92.11802	139321716	87.04531
Smp_2	184347011	166107090	90.10566	154134947	83.61131
Smp_3	156654062	142908461	91.22551	134425886	85.81066

Supplementary Table 5-2. Sequencing statistics for 96 cfDNA TAPS libraries.

Sample	Total reads	Uniquely mapped reads	Uniquely mapped deduplicated reads	% uniquely mapped reads	% uniquely mapped deduplicated reads
Cirrhosis_1	301693349	285664950	234084622	94.69	77.59
Cirrhosis_2	349158558	330739897	277565352	94.72	79.5
Cirrhosis_3	317186367	300714778	244731524	94.81	77.16
Cirrhosis_4	386827429	364916385	297474889	94.34	76.9
Ctrl_1	316209760	301601826	245756159	95.38	77.72
Ctrl_2	369523479	352249705	287183120	95.33	77.72
Ctrl_3	331431358	313730672	250202835	94.66	75.49
Ctrl_4	344841516	327821730	264086338	95.06	76.58
Ctrl_5	326582878	308673724	255925296	94.52	78.36
Ctrl_6	357295567	338532807	276867221	94.75	77.49
Ctrl_7	266641522	253903329	209446708	95.22	78.55
Ctrl_8	397006826	378187805	304074746	95.26	76.59
Ctrl_9	361760340	343179457	281211203	94.86	77.73
Ctrl_10	338796737	322171474	265283105	95.09	78.3
Ctrl_11	328322805	311096144	256679352	94.75	78.18
Ctrl_12	324145850	307735453	253637530	94.94	78.25
Ctrl_13	314877634	299816541	240140135	95.22	76.26
Ctrl_14	430135956	407928698	335424447	94.84	77.98
Ctrl_15	426406611	403985464	303902872	94.74	71.27
Ctrl_16	368705891	348948016	283719602	94.64	76.95
Ctrl_17	403785762	383599215	291607193	95	72.22
Ctrl_18	283439542	269291863	219596160	95.01	77.48
Ctrl_19	344779944	328946923	271368756	95.41	78.71
Ctrl_20	341460286	323592199	266662481	94.77	78.09
Ctrl_21	418080244	396340911	322976486	94.8	77.25
Ctrl_22	427486185	406722489	329856177	95.14	77.16
Ctrl_23	702676943	667388425	524487522	94.98	74.64
Ctrl_24	379922314	362597083	301273739	95.44	79.3
Ctrl_25	329363294	312809791	251449036	94.97	76.34
Ctrl_26	396526952	375662438	299045697	94.74	75.42

Ctrl_27	272564834	259886089	217463855	95.35	79.78
Ctrl_28	373281137	354535085	283643750	94.98	75.99
Ctrl_29	404843617	383003732	315083437	94.61	77.83
Ctrl_30	413728007	392646671	319704940	94.9	77.27
HCC_1	320638330	303621724	255140616	94.69	79.57
HCC_2	401904325	379506529	309638502	94.43	77.04
HCC_3	286990658	272119592	231043662	94.82	80.51
HCC_4	397769407	372615576	295311771	93.68	74.24
HCC_5	373386160	352598326	288006827	94.43	77.13
HCC_6	332555354	314468063	260893283	94.56	78.45
HCC_7	384206778	363417240	295151629	94.59	76.82
HCC_8	429728504	407254990	334620771	94.77	77.87
HCC_9	330041898	311558093	257186619	94.4	77.93
HCC_10	422057140	396481830	319836250	93.94	75.78
HCC_11	457976937	434422771	347903385	94.86	75.97
HCC_12	345005257	325946614	267210792	94.48	77.45
HCC_13	414644293	392076477	314532699	94.56	75.86
HCC_14	451163351	427561353	340679845	94.77	75.51
HCC_15	376688106	355687185	291604137	94.42	77.41
HCC_16	318568495	301931920	252252344	94.78	79.18
HCC_17	307658811	292869894	241188720	95.19	78.39
HCC_18	306492684	289872808	241932879	94.58	78.94
HCC_19	305904220	290457081	234831521	94.95	76.77
HCC_20	333128899	313208032	260667402	94.02	78.25
HCC_21	279349140	264140230	217538534	94.56	77.87
HCC_22	359608758	340779680	283832107	94.76	78.93
Pancreatitis_1	375130259	355689206	288811385	94.82	76.99
Pancreatitis_2	342235338	320085403	250902946	93.53	73.31
Pancreatitis_3	295480236	278163887	229061239	94.14	77.52
Pancreatitis_4	345848570	327543039	265527370	94.71	76.78
Pancreatitis_5	261268931	248512124	204826095	95.12	78.4
Pancreatitis_6	334331051	317177959	262199504	94.87	78.43
Pancreatitis_7	391140103	370243402	295945264	94.66	75.66
PDAC_1	325060871	309517181	257660062	95.22	79.27
PDAC_2	251683256	237813598	195186345	94.49	77.55
PDAC_3	374744744	356319496	283329483	95.08	75.61
PDAC_4	256917956	243383372	198361852	94.73	77.21
PDAC_5	384510247	365212239	294852731	94.98	76.68
PDAC_6	317535007	299517244	250304551	94.33	78.83
PDAC_7	380642398	362156005	291412890	95.14	76.56
PDAC_8	440991223	420097944	346022716	95.26	78.46
PDAC_9	418613483	400688829	304821023	95.72	72.82
PDAC_10	445110289	421596775	342363516	94.72	76.92
PDAC_11	365726839	346410429	285667634	94.72	78.11

PDAC_12	400810788	379728144	309024711	94.74	77.1
PDAC_13	348663643	332293525	269457090	95.3	77.28
PDAC_14	337532840	319037977	261039316	94.52	77.34
PDAC_15	358259137	339311549	279502330	94.71	78.02
PDAC_16	365231346	346731513	282586463	94.93	77.37
PDAC_17	412615619	391265495	315934380	94.83	76.57
PDAC_18	457020762	431975506	353265406	94.52	77.3
PDAC_19	301295703	285423508	230621613	94.73	76.54
PDAC_20	276761774	262905096	217008296	94.99	78.41
PDAC_21	340941884	324525639	238270088	95.19	69.89
PDAC_22	342697802	325059268	266440490	94.85	77.75
PDAC_23	282591403	268312613	224333787	94.95	79.38

Supplementary Table 5-3. Clinical characteristics of non-cancer control cohort.

Sample ID	Sex	Age	Comorbidities
Ctrl_1	Female	53	CIBH, Ulcerative Colitis
Ctrl_2	Male	71	CIBH, Hypertension
Ctrl_3	Female	66	CIBH, Asthma, Bronchiectasis, Polymyalgia, Coeliac
Ctrl_4	Female	60	Hypertension
Ctrl_5	Male	58	Rectal bleeding
Ctrl_6	Female	46	NA
Ctrl_7	Male	79	Rectal bleeding, CIBH, Encephalitis
Ctrl_8	Female	65	Chronic diarrhoea, Diabetes
Ctrl_9	Male	69	Surveillance polyps, Epilepsy, Previous polyps, Hypertension
Ctrl_10	Female	61	Hypertension
Ctrl_11	Male	48	IBDu, Ulcerative colitis
Ctrl_12	Male	69	Ulcerative Colitis
Ctrl_13	Female	45	Ulcerative Colitis, Distal proctitis
Ctrl_14	Female	45	TI focal active inflammation
Ctrl_15	Male	64	Crohn's disease, Possibly UC
Ctrl_16	Male	72	NA
Ctrl_17	Female	70	IBS, Sicca syndrome, Hypothyroidism
Ctrl_18	Female	68	PR bleeding
Ctrl_19	Female	65	Diabetes, Underactive thyroid, Hypertension
Ctrl_20	Male	78	AF, Hypertension
Ctrl_21	Male	50	Hypertension
Ctrl_22	Male	69	Duodenal polyp, Cardiomyopathy
Ctrl_23	Male	68	Duodenal polyp
Ctrl_24	Female	66	NA; part of bowel cancer screening program
Ctrl_25	Male	62	COPD, Hypertension
Ctrl_26	Male	81	IDA, Large prostate, Osteoarthritis, Hypopituitarism, TIA
Ctrl_27	Female	48	CIBH, IDA
Ctrl_28	Female	75	Osteoarthritis, Raynaud's, Diverticular disease, Polyp
Ctrl_29	Male	72	CABG, Aortic valve disease, Aortic valve replacement; LGD tubular adenoma in D3; atrophic gastritis
Ctrl_30	Female	75	IDA, Hypertension, Angina, LVE, Hemochromatosis, Hyperthyroidism

Supplementary Table 5-4. Clinical characteristics of PDAC cohort.

Sample ID	Sex	Age	TNM stage	Stage	Size of primary tumour (mm)
PDAC_1	Female	78	T2N2M0	III	28
PDAC_2	Male	60	T3N1M0	II	40
PDAC_3	Male	78	T3N2M0	III	45
PDAC_4	Male	75	T3N2M0	III	42
PDAC_5	Male	78	T3N2Mx	III	30
PDAC_6	Male	73	T3N1M1	IV	40
PDAC_7	Female	79	T3N1M0	II	36
PDAC_8	Female	59	T2N2M0	III	40
PDAC_9	Female	80	T3N1M0	II	60
PDAC_10	Male	62	T2N0M0	I	27
PDAC_11	Male	76	T2N2M0	III	30
PDAC_12	Female	76	T2N2M0	III	38
PDAC_13	Female	70	T2N1M0	II	30
PDAC_14	Male	71	T3N2M0	III	45
PDAC_15	Male	66	T2N1M0	II	60
PDAC_16	Female	74	T1N0M0	I	18
PDAC_17	Male	66	T2N2M0	III	30
PDAC_18	Male	67	T2N1M0	II	24
PDAC_19	Female	75	T1cN0M0	I	12
PDAC_20	Female	66	T2N1M0	II	40
PDAC_21	Male	71	T3N2M0	III	35
PDAC_22	Male	57	T2N1M0	II	35
PDAC_23	Male	69	T2N2M0	III	40

Supplementary Table 5-5. Clinical characteristics of HCC cohort.

Sample ID	Gender	Age	Cirrhosis	Aetiology	TNM stage	Stage	Tumour size (mm)
HCC_1	Male	75	No	No CLD	T3aN0M0	III	130
HCC_2	Male	62	Yes	ALD	T1N0M0	I	27
HCC_3	Male	64	Yes	ALD	T2NxM0	II	25
HCC_4	Male	57	Yes	NAFLD	T1N0M0	I	23
HCC_5	Male	78	No	None	T4NxM1	IV	160
HCC_6	Male	71	Yes	NAFLD	T3aN0M0	III	70
HCC_7	Male	78	Yes	NAFLD/ALD	T3aN0M0	III	87
HCC_8	Male	76	No	No CLD	T2N0M0	II	75
HCC_9	Male	73	Yes	NAFLD	T2NxM0	II	25
HCC_10	Male	69	No	NAFLD	T1NxM0	I	50
HCC_11	Female	72	No	PBC	T2N0M0	II	30
HCC_12	Male	69	Yes	MS	T2N0M0	II	17
HCC_13	Male	67	Yes	MS	T1N0M0	I	16
HCC_14	Male	71	Yes	Mixed ALD/MS	T1N0M0	I	12
HCC_15	Male	73	Yes	NAFLD	T2N0M0	II	28
HCC_16	Male	74	No	NAFLD	T1N0M0	I	54
HCC_18	Male	85	No	No clear CLD	T3aN0M0	III	118
HCC_19	Male	62	Yes	MS	T3bN0M0	III	30
HCC_20	Male	62	Yes	ALD	T1N0M0	I	9
HCC_21	Male	67	Yes	NAFLD	T3aNxM0	III	33
HCC_22	Male	89	Yes	Cryptogenic	T1N0M0	I	24

Supplementary Table 5-6. Clinical characteristics of pancreatitis cohort.

Sample ID	Sex	Age	Diagnosis	Comorbidities
Pancreatitis_1	F	81	Chronic pancreatitis with cysts	Diabetes
Pancreatitis_2	M	53	Chronic calcific pancreatitis	Diabetes
Pancreatitis_3	M	86	IgG4-autoimmune pancreatitis and systemic disease (SMG. Lungs)	Diverticular disease, interstitial lung disease
Pancreatitis_4	M	49	IgG4- burnt out pancreatic disease with retroperitoneal fibrosis	Anxiety
Pancreatitis_5	F	35	Acute on Chronic pancreatitis with pseudocysts	Depression
Pancreatitis_6	F	35	Hereditary chronic pancreatitis	Osteopenia
Pancreatitis_7	M	37	Recurrent idiopathic pancreatitis	Diverticulosis, pyloric stenosis

Supplementary Table 5-7. Clinical characteristics of cirrhosis cohort.

Patient ID	Sex	Age	Status	Aetiology
Cirrhosis_1	Male	40	Cirrhosis	ALD
Cirrhosis_2	Female	61	Cirrhosis	ALD
Cirrhosis_3	Male	74	Cirrhosis	ALD
Cirrhosis_4	Male	45	Cirrhosis	ALD

Supplementary Table 5-8. Summary of performance of prediction models trained with various genomic features. The best achieved AUC are shown in the table.

Genomic features	HCC - AUC	PDAC - AUC
DMC	0.88	0.81
Protein coding	0.78	0.91
SINE	0.88	0.78
LINE	0.85	0.78
Promoter	0.87	0.96
CGI shore	0.84	0.73
CGI	0.97	0.70
LTR	0.82	0.74
CGI shelf	0.86	0.87
Enhancer	0.99	0.76
Open chromatin	0.89	0.80
TF binding site	0.84	0.82

10.6 Appendix 6 – Supplementary Materials to Chapter 6

Supplementary Table 6-1. Mapping and sequencing quality statistics for eeTAPS libraries.

Library	eeTAPS (200 ng)	eeTAPS (1ng)	eeTAPS (10ng)	eeTAPS (50ng)	eeTAPS (200 ng, rep2)	eeTAPS (ctrl)
Total reads	75087691	25488766	22371125	28370411	107627064	9568590
Mapped reads	74487124	25433543	22327287	28315992	107376184	9546719
Properly mapped reads	67849413	15727887	20824202	27475285	103168329	9232041
% mapped reads	99.2	99.7	99.8	99.8	99.7	99.7
% properly reads	90.3	61.7	93	96.8	95.8	96.4
Sequencing depth	4.53	1.54	1.35	1.71	6.49	0.58

Supplementary Table 6-2. Mapping and quality statistics for wgTAPS and rrTAPS.

	wgTAPS	rrTAPS
Total reads	455548210	41956461
Mapped reads	406057179	37017706
Properly mapped reads	399138623	36401314
% mapped reads	89.1%	88.2%
% properly reads	87.6%	86.0%
Sequencing depth	27.45391443	2.5285339