DNA methylation in late-stage oesophageal cancer

Phil Xie

Lincoln College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2023

This thesis is dedicated to

All curious minds, who
are amazed by the palindromic nature of CpG

```
CATAATATACGCGCTGCTTTCGTTCGCCAC
     GCGC        GC  GC
```

and my dear friends and family
whom I am proud of

. . . and hopefully proud of me?

Preface

How to read the PDF

This is relevant to readers who are viewing the document electronically. Most references and cross-references to figures or text should have a hyperlink, which the readers can click and bring them to the linked material.

To return to the original position, most PDF viewers have a "Back" function, which can usually be accessed via the shortcut `Alt + left arrow` or `Cmd + left arrow`. However, each application may have a different shortcut, and the readers may have to check their viewers' shortcut instructions.

Regarding Chapter 3

Chapter 3 may be rather unfriendly to those who are unfamiliar with statistical concepts and mathematical notations. In the best interest of readers, I have shaded the technical aspects of the methodology in grey. While I would strongly encourage readers to critically appraise the shaded materials, the details are not required for comprehending the rest of the thesis, as long as a high-level conceptual understanding is obtained.

Acknowledgements

Personal

My envisioned timeline in the beginning of my DPhil was initially a balanced combination of wet and dry lab - sample processing and data collection, data analysis of unbiased molecular profiling to generate hypothesis, and finally going back to the bench to validate the hypotheses. However, the bioinformatic analysis of bulk quantitative assay is more challenging than I would have possibly imagined. The complex biology of DNA methylation is another big hurdle. Hence, this thesis ended being mostly computational.

I would like to start by thanking Xin Lu, my primary supervisor, for trusting me with this project and giving me the autonomy to explore the project along with my research interest. She has the amazing ability to see the silver lining in any disappointing situation I am in, and can always point me to the right person when I needed help in specific expertise. I would also like to thank Benjamin Schuster-Böckler, who has been the *de facto* supervisor in my bioinformatic analyses. I graduated from medical school without statistical or computational background, and Ben has been instrumental in my journey into bioinformatics. Furthermore, I also need to thank Skirmantas Kriaucionis and Chunxiao Song for their kind comments and much needed discussions in the project, including but not limited to the biology of DNA methylation, techniques in measuring it, and possible research directions. I am extremely grateful to have them in my supervisory team.

I need to wholeheartedly thank my partner, Vanessa. We supported each other throughout COVID and doctorate studies, and shared our ups and downs in life. It is also because of her that I dug deep into the basics of regression models and linear algebra, without which this thesis would have been impossible.

I need to thank my friends in science, who supported me along the way. Nick has been a great friend and inspiration for me, although our research field barely overlaps, I learnt from his incredible attitude to science and I genuinely hope that I will have similar achievements as he does. Andrew is my best buddy who has an almost identical career path as I do, and his conversations with me kept my passion burning. Paulina and Masato were DPhil students from the Song lab, and I learnt everything I knew about TAPS and library preparation from them. Each batch of TAPS sequencing is painstakingly executed, and I think we could start

a religious cult around pipetting techniques. Magda is my buddy also working on the bioinformatic analyses of the Trial project. We shared the highs and lows of DPhil, sent memes to each other, and also tips and tricks in using R. As mentioned, my background in maths and statistics is weak, and whenever I have problems I would discuss with Terence and Otto, who are real mathematicians.

I also need to thank Angie Green, who is an account manager at the former Oxford Genomics Centre. I have never met her in person, but I literally learnt everything about library pooling, repooling, and index compatibility from her. I would also like to thank the stranger who would always park his car a little away from the bike rack, just so I could get my bike out easier. This small act of kindness has brightened my every morning.

Last but not least, my beloved Mom and Dad. I made it, I made it to the end. I still cannot believe that I am now at my parents' age when I was born. Thank you for taking care of me and nurturing my passion for the many years that have passed. James, thank you for being a great brother and showing me what it is like to pursue one's own interest. If not for my family, I would not even have the chance to become interested in science, having grown up in a highly capitalist society where socioeconomic status is the core pursuit for most people.

Institutional

The Oxford University has provided an excellent environment for research and organized many useful skill training courses. In particular, I would like to thank the organizers and tutors of the Oxford Biomedical Data Science training courses, which helped me to build a solid foundation in the pursuit of data science research.

I would like to thank the Faculty of Medicine of The Chinese University of Hong Kong, where I studied as an undergraduate. Unlike in Oxford, it was unusual for medical school students to take part in research in Hong Kong, but the Faculty facilitated this by pioneering the Global Physician Programme, and allowed me to accumulate research experiences to make the transition for DPhil studies. I also benefitted from the Mabel Churn Scholarship for postgraduate studies.

I need to thank the Croucher Foundation for their generous stipend during my DPhil. The academic allowance fund helped me to attend courses and are beneficial for both my current studies and future research interest.

Finally, I would like to thank the Ludwig Institute of Cancer Research, which is my home institute in Oxford. The Institute funded part of my studies, organized extremely useful student activities and eye-opening seminars, and maintained a highly organized and interdisciplinary research environment. Most importantly, my thesis project would not have happened if not for the Institute.

Abstract

Oesophageal adenocarcinoma (OAC) is a cancer with unmet needs, signified by late diagnosis, poor prognosis, and ineffective screening. Recently, phase III randomized trials showed that first line combined immune checkpoint inhibitors (ICI) and chemotherapy (CTX) is superior to CTX alone in selected patients. However, patient selection is based on histologic criteria that could vary substantially between paraffin sections. Also, there is a lack of good tools for ICI treatment monitoring.

It has been reported that changes in DNA methylation are common and early events in the development of OAC. DNA methylation can also be peripherally detected in circulating cell-free DNA (cfDNA), which showed promises in early detection of cancer and disease monitoring. Therefore, I investigated whether DNA methylation is associated with prognosis in OAC after ICI+CTX, and whether tumour-specific signals can be detected in cfDNA methylation.

The LUD2015-005 study is a phase I/II trial to assess the safety and efficacy of combined ICI+CTX in OAC. Samples from 23 patients with at least 12 months of follow-up belonging to the inoperable cohort were used for whole genome sequencing (WGS), bulk RNA sequencing, and whole genome single-base resolution methylation using TET-assisted pyridine borane sequencing (TAPS).

A new statistical framework was proposed to conduct tumour purity and copy number aware analysis of the tumour DNA methylation. Global hypomethylation in late replicating regions was identified to a specific feature for all OAC and possibly Barrett's oesophagus, the pre-malignant lesion of OAC. Importantly, the tumour-specific methylation can be detected in cfDNA down to around 0.2% tumour fraction, and changes in cfDNA tumour signal after 4 weeks of ICI can predict progression-free survival up to 1 year.

In addition, two OAC methylation subtypes were discovered, which were characterized by genomic instability and severe global hypomethylation, versus less severe global hypomethylation but more aggressive molecular phenotype. The clinical implication of these subtypes is currently unclear due to small sample size.

Finally, this thesis highlighted a possible mechanism of focal hypermethylation in cancer by linking together the crosstalk between replication timing, histone modification, and DNA methylation, which is yet to be tested with further experiments.

List of frequent abbreviations

5mC 5-methylcytosine. 15

ATAC-seq Assay for Transposase-Accessible Chromatin with sequencing. 8

BER base excision repair. 18, 19, 29
BO Barrett's oesophagus. 2, 6–9, 11–13, 28–30, 80

CB clinical benefit. 33
cfDNA cell-free DNA. 4, 13, 14, 32, 36, 51, 82
CGI CpG island. 21, 22, 47
CIMP CpG island methylator phenotype. 55
CNA copy number alteration. 37, 38
CPS combined positive score. 2
ctDNA circulating tumour DNA. 14
CTX chemotherapy. 2, 4, 31–33

DEG differentially expressed gene. 103
DMR differentially methylated region. 41
DNMT DNA methyltransferase. 73
DSB double strand break. 19, 20
DTA diphtheria toxin fragment A. 9

GOJ gastroesophageal junction. 7, 8, 11, 12, 34
GORD gastroesophageal reflux disease. 13, 80
GSEA gene set enrichment analysis. 103

hyperDMR differentially hypermethylated regions. 49, 70–73, 76, 79, 81
hyperT less-hypomethylated tumour. 95
hypoDMR differentially hypomethylated regions. 49, 67–69, 79, 81
hypoT hypomethylated tumour. 95

ICI immune checkpoint inhibitor. 2–4, 31–33
ICR imprinting control region. 22
irCR complete response. 33
irPD progressive disease. 33

irPR partial response. 33
irSD stable disease. 33

L2FC log2 fold change. 75, 103
LOH loss of heterozygosity. 29, 37
LTSR long term survivor recall. 33

MeDIP methylation analysis by DNA immunoprecipitation. 26
MLE multilayered epithelium. 9, 11
MMR mismatch repair. 3, 29, 30
MSI microsatellite instability. 37

NCB no clinical benefit. 33
NMF non-negative matrix factorization. 30

OAC oesophageal adenocarcinoma. 1, 2, 5, 7, 8, 11–13, 28–31
OR odds ratio. 68
OS overall survival. 33
OSCC oesophageal squamous cell carcinoma. 5, 30, 80
OSMG oesophageal submucosal gland. 10–12

PCR polymerase chain reaction. 14
PFS progression-free survival. 33
PMD partially methylated domain. 26, 27, 30
PostTx post combined immunochemotherapy treatment. 33
PRC Polycomb repressive complex. 73
PreTx pre-treatment baseline. 33

RE repeat element. 67–69
REC residual embryonic cell. 9, 10, 12
RepD replicating domain. 67–73
RNA-seq RNA sequencing. 4, 34, 36, 37

scRNA-seq single cell RNA sequencing. 8, 11, 12, 31
SNV single nucleotide variant. 8

TAPS TET-assisted pyridine borane sequencing. 35–38
TET Ten-Eleven Translocation. 73
TMB tumour mutational burden. 2, 3
TMD tumour methylation detected. 83–85
TSG tumour suppressor gene. 93
TSS transcription start site. 71

WES whole exome sequencing. 3, 99
WGD whole genome duplication. 37
WGS whole genome sequencing. 3, 4, 37

Contents

List of Figures

1 Introduction

Contents

1.1 Motivation and context

Epigenetic reprogramming has emerged as a new hallmark of cancer [1]. Mutations in epigenetic regulators as well as epigenetic silencing of tumour suppressor genes has been reported across multiple tumour types. Oesophageal adenocarcinoma (OAC) is one of such tumours that has been reported to harbour epigenetic changes

early on, as early as in its pre-malignant stage of Barrett's oesophagus (BO) [2–8]. One of the most well-known affected gene in OAC is the hypermethylation of CDKN2A (or p16INK4a), a tumour suppressor gene which was thought to be silenced via the epigenetic mechanism.

The context of this study happens in a time when there is an explosion of cancer immunotherapy researches and clinical trials using immune checkpoint inhibitors (ICIs). ICI has almost become a miracle drug, being able to treat a myriad of cancer types [9], and it is not unheard of that even metastatic patients can attain durable complete response after receiving immunotherapy [10]. However, despite the huge success in ICI, only a subset of patients benefits from the treatment, and it remains a challenge to accurately identify this population. Exemplars of predictive biomarkers for ICI response include PD-1 or PD-L1 expression and tumour mutational burden (TMB), but both biomarkers have their caveats.

1.1.1 ICI in late-stage OAC

Recently, first line combined ICI and chemotherapy (ICI+CTX) has been approved by the United States Food and Drug Administration (FDA) in treating oesophageal and gastroesophageal cancers [11]. As of the time of writing, current guidelines recommend first line ICI+CTX for patients with a combined positive score (CPS) of ≥ 10 based on phase III randomized clinical trials [12–14], with some centres recommending a CPS cutoff of ≥ 5. CPS is a scoring method for PD-L1, defined as the total number of PD-L1 stained cells, including both tumour and immune cells, divided by the total number of viable tumour cells and multiplied by 100. Compared to other PD-L1 scoring methods, CPS was shown to have excellent inter-pathologist agreement [15]. However, one must note that CPS is calculated based on only one or few histology sections. While the score has a small technical variation, I cannot help but wonder how much biological fluctuations there would be if more slides were sectioned, in the context of intratumoural heterogeneity as well as variable biopsy quality.

TMB is a scoring method for the frequency of mutations in tumour cells, with slightly different definitions in different studies. As of now, the standard way of reporting TMB is the number of non-synonymous mutations per million base-pair (Mb) genome sequenced. There is typically a remarkable response to ICI for patients with mismatch repair (MMR) deficiency [16], which results in extraordinarily high TMB. In cases with no MMR deficiency, a high TMB also generally associates with better ICI outcomes [17], but may not be predictive in certain cancer types such as renal cell carcinoma [18]. The problem with TMB is that currently there is no universal standard of how it should be performed. It can be measured by many platforms, such as whole genome sequencing (WGS), whole exome sequencing (WES), targeted gene panels, and other proprietary assays. The measured TMB using different methods may not be equivalent, due to differences in panel size or sequenced genome, sequencing depth, use of paired germline references, and sample processing protocol. The issue with TMB is particularly significant when the biopsy has a low tumour cell content (referred to in the field as "tumour purity"), where TMB would be artefactually low due to decreased chance of detecting somatic variants [19–21].

1.1.2 Aims of study

It is very likely that no single biomarker is perfect, given the complicated nature of tumour biology and complex immune microenvironment. I believe the key to precision treatment is to have multiple high confidence, independent biomarkers. As reviewed in [22], epigenetic changes may play a role in ICI. However, much of the focus seems to be on the epigenetic modulation of selected genes, and as a "bulk" biomarker without considering the cellular contribution of the epigenetic mark, for example in [23]. This is understandable as the technology for cell-resolution or spatial epigenetics is still under development, but also problematic because unlike somatic mutations, epigenetic changes can happen in both tumour and normal cells. For example, epigenetic silencing of proliferation-related genes in tumour versus immune cells can have opposite implications clinically.

It is from this angle that I approached the LUD2015-005 trial [24], which is an open-label phase I/II study to evaluate the safety and efficacy of combined ICI+CTX in lower oesophageal cancer. Each patient is deeply profiled for multi-omics data, including WGS, whole methylome sequencing, and RNA sequencing (RNA-seq). Longitudinal tumour biopsies were taken via endoscopy throughout treatment, and circulating cell-free DNA (cfDNA) were also sequenced for DNA methylation at multiple timepoints. With the wealth of parallel information, I took on the challenge of (1) analysing bulk DNA methylation data in the context of a mixture of cell types, and (2) obtaining clinically meaningful information from DNA methylation. I would also like to use the privilege of having this data to (3) generate testable hypotheses about fundamental tumour biology from clinical samples.

1.2 Introduction to oesophageal adenocarcinoma

Oesophagus is a tubular organ with one single function, which is to allow substances to pass from the oral cavity into the stomach. Whatever food an animal swallows, hot or cold, soft or hard, dead or alive, it will pass down the oesophagus. The same goes for whatever is being vomited. Together with acidic gastric juice or alkaline bile secretion, it will travel up the oesophagus.

The oesophagus also has a rather awkward anatomical position. For whatever reason, evolution decided that it is a good idea to bundle this passage for potentially hazardous materials together with trachea, heart, and aorta, all of which would lead to dire consequences if damaged. Personally, I am very convinced that this is the reason why humans share a common fear of ingesting things alive, as it would be very unfortunate had the oesophagus been perforated.

There are two major symptoms when things go wrong in the oesophagus, related to the function and anatomical position respectively. One is dysphagia, which means difficulty in swallowing. The other is chest pain, which can come in different descriptions, such as burning, squeezing, stabbing, or even crushing pain. However, sinister oesophageal diseases such as cancer are typically asymptomatic until later on, when the tumour is big enough to obstruct the food pipe, or has already

invaded other organs. In this section, I will review the clinical needs and origins of oesophageal adenocarcinoma.

1.2.1 Survival of patients with oesophageal cancer

Perhaps due to the late symptom onset and the proximity to vital organs, oesophageal cancer is often already unresectable upon diagnosis. According to analysis of data released by public health authorities in the United Kingdom (UK) [25–28], oesophageal cancer have the second-highest proportion of late-stage disease (stage III/IV) at initial diagnosis (Fig. 1.1a) among common cancers, just after pancreatic cancer. Late-stage disease is often associated with worse survival outcome. According to data from England [29] between 2016 and 2020, oesophageal cancer has one of the worst survival (Fig. 1.1b), together with pancreatic and lung cancers, all of which have more than 70% of late-stage disease at diagnosis. Even for early-stage resectable oesophageal cancer, patients have to undergo major operation involving the thoracic area, which require expertise from multiple surgical specialties, and is associated with significant risks and post-operative complications.

Note that the statistics presented are for all oesophageal cancers combined, which consists of OAC and oesophageal squamous cell carcinoma (OSCC). The histological subtype of interest in this thesis is OAC.

1.2.2 Origin(s) of OAC and Barrett's

The oesophagus is lined with stratified squamous epithelium, a specialized epithelium that contains high amount of cellular tight junctions, which allow it to withstand mechanical abrasions and stress. It makes sense that OSCC can arise from the squamous epithelium, but where does OAC, a glandular type of cancer, originates from?

The history of this debate is surprisingly recent [30], as reviewed in detail by Spechler et al. Briefly, in the early 1950s, Norman Rupert Barrett and a couple other investigators reinvestigated the presence of ulcers accompanied by the histological findings of columnar epithelium in the lower oesophagus. This

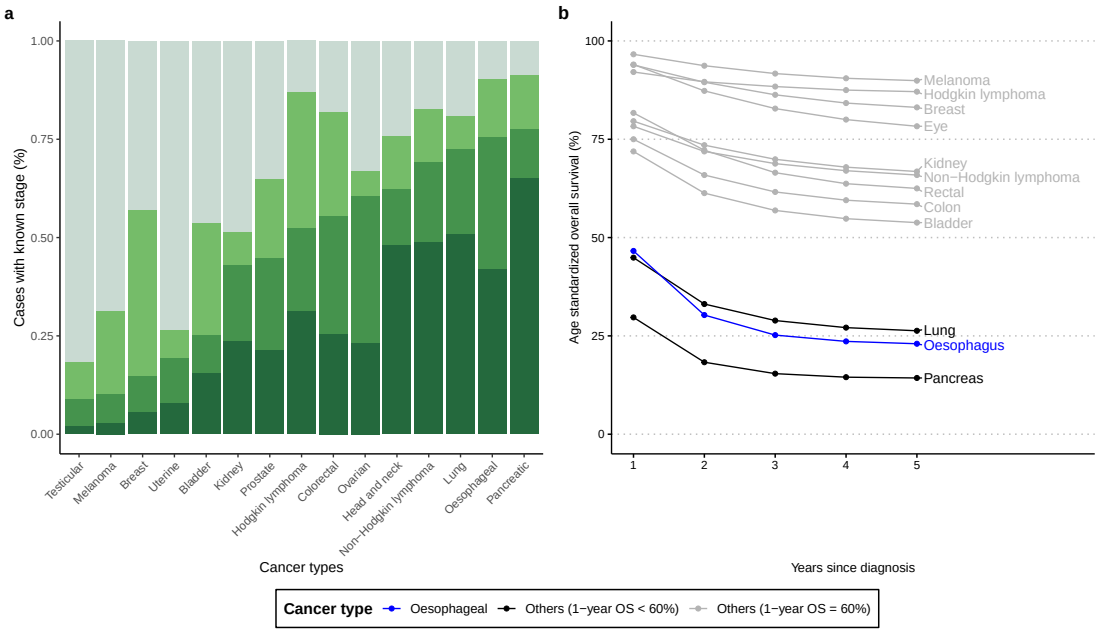


Fig. 1.1: Stage at diagnosis and survival trend of common cancers. (**a**) Top 15 common cancers in UK ordered by proportion of stage III and IV disease at diagnosis. Oesophageal cancer has the second-highest proportion of late-stage diagnoses, and fifth-highest proportion of stage IV diagnoses. (**b**) Age-standardized overall survival of common cancers in England. Lung, oesophageal, and pancreatic cancers have the lowest 1-year survival.

lesion is subsequently named after Dr. Barrett, and is today known as the Barrett's oesophagus. BO was a rare diagnosis back then, and the reported cases often had coexisting hiatal hernia. Owing to that, BO was initially thought to be an upward displacement of the entire stomach as an organ, which is lined by columnar epithelium [31]. Later on, investigators provided more evidence that it was only the epithelium that was gastric-like [32, 33], whereas the deeper structures are those typical of an oesophagus, including the presence of oesophageal submucosal glands and the gross appearance of a tubular structure. It was hence increasingly recognized that BO is actually columnar metaplasia of the squamous oesophageal epithelium, instead of being part of stomach.

The tumorigenic potential of BO was established in the next few decades. A retrospective study by Haggitt et al. [34], reviewed the historical specimens of adenocarcinomas involving the oesophagus in a single centre. Over a period of 50

years, 14 cases of OAC were identified, out of which 12 were found to arise in a background of BO, and 1 is too ulcerated to be assessed. At this point, it was widely accepted that BO is the origin of OAC, although the hypothesis was revisited 40 years later, where a study found that about half of OAC patients were not found to have of BO [35]. It was not a fair comparison though, because the diagnostic criteria of BO has evolved over time, and in the more recent study, diagnosis of BO required the presence of not only columnar-lined epithelium, but also intestinal metaplasia. With the advances in sequencing, numerous molecular similarities including shared somatic mutations and epigenetic features were found between BO and OAC [2–8, 36], supporting the notion that BO is a pre-invasive form of OAC. As of the time of writing, the current paradigm remains that OAC likely originates from BO. Howbeit, this is not even close to the end of the discussion, and brings us to our next question - what then, is the origin of BO?

Determining the cell-of-origin of BO has been particularly tricky, partly because it is typically found near the gastroesophageal junction (GOJ), where two specialized epithelium meet with vastly different cell types, which led to many possible hypotheses, as thoroughly reviewed by Que et al. [37]. The main hypotheses include proximal migration of gastric cardia cells, differentiation of KRT7^{high}stem cells at the GOJ, and differentiation of stem cells in oesophageal submucosal glands and ducts. The evidence for each hypothesis will be discussed in the following paragraphs. Another reason why origin of BO (and OAC) is still a debate is the lack of convincing animal models [38]. In particular, the relevance of using mouse models in the study of human oesophageal pathology is unknown. Human oesophagi are non-keratinized and have submucosal glands, whereas rodents have keratinized oesophagi without submucosal glands. Rodents also have a forestomach that is lined with squamous epithelium, so there is in fact no transition of cell types at the anatomical GOJ.

Gastric cardia cells

The human stomach is not a uniform organ, but rather divided from rostral to caudal into the cardia, body, and antrum, with each part lined by slightly different

mucosae. Gastric cardia is the part in direct contact with the oesophagus and GOJ, and comprised mainly mucous-secreting cell types. There is profound similarity between the histology of cardiac mucosa and BO, and often the presence of goblet cells or submucosal tissue is needed to make a definitive diagnosis of BO [39].

Quante et al. [40] provided experimental evidence for this hypothesis. Instead of the commonly used surgical reflux model, the authors induced chronic inflammation in the squamous oesophagus and forestomach by overexpression of IL-1β. Using Cre-loxP mediated lineage tracing, they found that Lgr$^+$ progenitor cells in the gastric cardia could migrate to the inflamed oesophagus and form BO-like metaplasic lesions after treatment with bile acids in the IL-1β transgenic mice.

Nowicki-Osuch et al. investigated the hypothesis in human and performed comprehensive molecular profiling of BO, normal oesophagus, stomach, duodenum, and oesophageal submucosal glands [36]. The analysis results of single cell RNA sequencing (scRNA-seq), methylation array, and Assay for Transposase-Accessible Chromatin with sequencing (ATAC-seq) all showed consistently that BO is most closely related to gastric cell types. Apart from correlative data, the authors also performed lineage tracing using clonal somatic single nucleotide variants (SNVs) to show that clonal SNVs in BO overlap with those in gastric, but never with oesophageal samples. However, concerns regarding the methodology of the lineage tracing have been raised [41], which may involve post-hoc selection of favourable cutoffs and potential false positive clonal SNVs in gastric samples.

KRT7$^{\text{high}}$ stem cells at the GOJ

KRT7 (or CK7) is one of the specific immunohistological markers used by pathologist to aid the diagnosis of OAC, and is strongly expressed in both BO and OAC [42]. Normally, KRT7 is expressed in certain secretary cells, and not or lowly expressed in oesophageal and gastric epithelium.

Wang et al. [43] described a possible source of KRT7$^{\text{high}}$cells. The authors first established that p63 is required for the development of squamous oesophagus and forestomach in mice. In p63 null mice, the oesophageal epithelium is lined by Krt7$^+$

columnar cells that resembles BO. The authors then showed that small amount of Krt7$^+$ cells are present at the squamocolumnar junction in adult wild-type mice. Because of the embryonic origin, these cells were named residual embryonic cells (RECs). Wang et al. then expressed diphtheria toxin fragment A (DTA) in squamous cells, creating an animal model that mimics epithelial injury, and showed that the squamous epithelium was replaced by Krt7$^+$ cells. No genetic lineage tracing technique was used, but histological sections showed that the Krt7$^+$ cells are contiguous with RECs at the squamocolumnar junction, implying that RECs may expand and migrate to repair the damaged epithelium.

Jiang et al. [44] described another possible source of KRT7^{high}cells. They first noticed that upon overexpression of SOX2, an important morphogen for foregut development [45], there was basal cell hyperplasia at the squamocolumnar junction, which is at the same location with the aforementioned RECs. To demonstrate the pathological relevance, oesophageo-gastroduodenal anastomosis surgery was performed to cause bile acid reflux, and led to the expansion of the same cells. The hyperplastic epithelium phenotypically resembles multilayered epithelium (MLE), which pathologists thought is a precursor of BO and pathognomonic of the condition [39, 46]. Unlike RECs, however, these basal cells are also p63$^+$, and have classical markers of both squamous and glandular cells. They are thus named "transitional basal cells". The authors further showed that overexpressing CDX2 in mice resulted in BO-like epithelium via expansion of the p63$^+$KRT7$^+$ cells. Importantly, numerous goblet cells appeared within the BO-like epithelium by week 13 of CDX2 overexpression, making this work the first BO animal model that recapitulates intestinal metaplasia. Jiang et al. also managed to generate organoid cultures from human transitional basal cells. Overexpressing CDX2 in human organoid cultures did not demonstrate goblet cell differentiation within the culture period, although expression of intestinal marker genes did increase.

The results from Wang and Jiang et al. may seem contradicting, as the former claimed that BO originates from p63$^-$KRT7$^+$ cells, whereas the latter proposed that BO originates from p63$^+$KRT7$^+$ cells. However, the two hypotheses can be

unified if the so-called RECs are actually derived from transitional basal cells. It is well known that p63 is only highly expressed in the basal layer of a stratified epithelium, and is lost as the cells move towards the luminal layer. Here I performed immunohistochemistry staining of p63 and KRT7 on a 15-week human foetal oesophagus to show that the findings in mice may be relevant in human as well. The entire oesophagus is lined by multilayered cells, with $p63^+$ basal layer and $KRT7^+$ luminal layer. Representative sections are shown in Fig. 1.2. This preliminary data supports the notion that the transitional basal cells described by Jiang et al. is the progenitor for RECs described by Wang et al., and both are remnants of an epithelial type that is prominent during embryonic development.

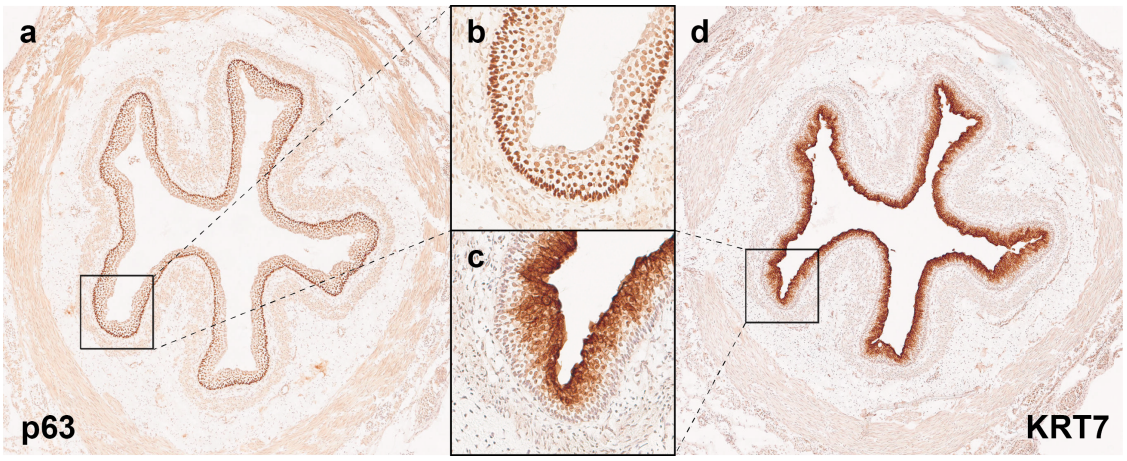


Fig. 1.2: Immunohistochemistry staining of p63 and KRT7 in 15-week human foetal oesophagus. Fresh abortus samples are collected via the Human Developmental Biology Resource (HDBR), and 15 µm cryosections are performed with staining. (**a,b**) p63 staining is restricted to the basal layer of the stratified epithelium. (**c,d**) KRT7 staining is restricted to the luminal layer of the stratified epithelium.

Oesophageal submucosal glands and ducts

The oesophageal submucosal glands (OSMGs) in humans are mucous glands situated beneath the muscularis mucosae, whereas in other mammals they often have a serous component [47]. The specific function of the glands remains unknown. Interestingly, Goetsch [47] noted that the presence of these glands may be related to the animals' diet, where animals that eat coarse food such as herbivores tend to have no OSMG, whereas those that eat soft food such as carnivores tend to have

OSMG. The topographical distribution of OSMG in humans seems to be highly variable across individuals [47]. A comprehensive survey of human OSMG was performed by van Nieuwenhove et al. [48], and found that OSMG is most abundant in the upper-third of the oesophagus (1.68% of total surface area), followed by lower-third (1%), and rare in the middle-third (0.03%).

Submucosal gland as an origin of cancer was described as early as 1962 [49], although it is likely that the authors were describing a different histological subtype than BO-derived OAC. Animal studies using canine models [50, 51] demonstrated the regenerative and metaplastic potential of OSMG ducts. In 2005, a study by Coad et al. [52] meticulously studied the morphological relationship between OSMG and BO in human, and found a striking evidence where there was a gradual transition of cell types from OSMG duct into BO, and captured the presence of MLE between the two structures. Leedham et al. [53] later on studied BO on the level of individual crypts, and found a p16 (CDKN2A) mutation shared between OSMG duct and the contiguous BO, providing lineage tracing evidence in human. This is further supported by the work of Nicholson et al. [54], who employed an innovative method of using functional defect of a mitochondrial enzyme CCO for lineage tracing. Albeit the authors were unable to find direct evidence of shared mutations between BO and OSMG, they were able to find shared mutations between BO and adjacent normal squamous epithelium, suggesting that the two epithelia were derived from the same stem cells. OSMG is known to have regenerative capacity for squamous epithelium [52, 53], and hence their results indirectly support OSMG giving rise to BO. Recently, Owen et al. [55] showed that BO is transcriptionally more similar to OSMG than gastric cells using deeply sequenced scRNA-seq. They also found expression of goblet cell markers in OSMG in patients with BO, providing evidence that OSMG can potentially undergo intestinal metaplasia.

Last but not least, there have been numerous independent case reports of the development of BO after total gastrectomy [56–66]. Since OSMG is the only structure that does not involve GOJ and stomach among the contested cells-of-origins, these reports are supportive of OSMG being the origin of BO.

Verdict

The origin of BO is still currently being contested, with each hypothesis being backed up by solid observational and experimental evidences. In my view, none of these evidences disprove other hypotheses, and it is most likely that BO can arise from multiple locations, although OSMG seems to have the best evidence in human.

As explained above, RECs and transitional basal cells are likely closely related entities, both with an embryonic origin. Development of OSMG is likely postnatal [67], but like the transitional epithelium at squamocolumnar junction, OSMG ducts also have a $p63^+$ basal layer and $KRT7^+$ luminal layer [68]. It has also been recently proposed through scRNA-seq data that OSMG and the $KRT7^{high}$population at GOJ may share the same developmental lineage [36]. Given the phenotypical and transcriptional similarities, it is not unacceptable that all these cell types can develop into BO.

1.2.3 Surveillance and early detection of OAC

One of the motivations to study the cell-of-origin of OAC and BO is to better inform the screening strategy of OAC. For example, OSMG is a deep structure that is not often picked up by endoscopic biopsy. If it is indeed one of the origins of BO, then biopsy strategies may need to be revisited.

Given strong evidence that BO is a risk factor for OAC, patients with a diagnosis of BO are often subject to surveillance endoscopy. In patients who were diagnosed of OAC, those who underwent surveillance endoscopy had earlier stage disease, as well as improve survival [69–71] compared to those without surveillance. However, these evidences were based on retrospective cohorts, and do not represent the absolute benefit of surveillance endoscopy in BO. Multiple large-scale population based studies have revealed that the absolute annual risk of progression from BO to OAC is quite low, probably between 0.1 to 0.4% [72, 73]. This raised questions about the design of surveillance programmes to maintain cost-efficacy, especially for BO without high risk features. BOSS, a 10-year prospective non-inferiority randomized trial, is currently being conducted for non-dysplastic or low-grade

dysplasia BO to evaluate whether regular surveillance is superior to as-needed endoscopy in terms of patients' overall survival and cost-effectiveness [74]. There has also been effort on identifying high risk BO using molecular characteristics such as aneuploidy [75] to inform patient stratification.

While how best to survey patients with known diagnosis of BO is an area of active research, the bigger problem is probably how to effectively identify patients with BO in the first place, or other important conditions that may be high risk for OAC. As highlighted by studies in the early 2000s, less than 5% of patients with OAC had a diagnosis of BO 6 months prior to their cancer [76, 77]. Also, in cancers with effective screening, it is expected more cancers would be diagnosed at an earlier stage, such as breast and colon cancers. Given the proportion of late stage diagnoses in oesophageal cancers (Fig. 1.1b), it is unlikely that the current strategy of OAC detection is effective.

Currently, BO remains an endoscopic diagnosis. Only when the endoscopist sees gross changes suspicious of BO, biopsy is taken for pathological confirmation. Endoscopy is not without risk, and therefore only performed in patients with appropriate indications, such as having chronic gastroesophageal reflux disease (GORD) and multiple risk factors [39]. In addition, the procedure has to be operated by specialist teams with expensive optics and equipments, and thus limited in throughput. Hence, the Cytosponge studies led by Fitzgerald et al. [78–80] is a welcoming one with high clinical value, bypassing the need of endoscopy to screen for BO. Cytosponge, fittingly, is a sponge-like cell collection device that can be administered by trained nurses, which scrapes the epithelium of the stomach and oesophagus for tiny pieces of tissue. The tissue fragments would then be processed for histology and TFF3 staining, a verified marker for BO with intestinal metaplasia [81]. In a randomized controlled trial, Cytosponge was administered to patients with GORD symptoms, and shown to diagnose about 10 times more BO compared to standard care [80].

Another attractive early detection strategy is liquid biopsy, for example circulating cfDNA. "Liquid biopsy" is a trendy umbrella term that was coined in

the early 2010s, initially referring to circulating tumour cells [82], and later on expanded to include all kinds of liquid-based molecular or biochemical assays [83]. Compared to diagnosing or characterizing tumours with traditional biopsy, liquid biopsy makes use of easily accessible body fluids such as blood, which makes it less risky, logistically more feasible, and cheaper to be performed. Regardless of which assay, a common theme among these tests is that the actual tumour-derived material is often exceedingly rare, which raises challenge in both the experimental assay design and analytical identification of tumour signals from the background noise. Among which, nucleic acid based assays have a unique advantage over other methods, because nucleic acids can be amplified using polymerase chain reaction (PCR). Although PCR induces bias and errors, this crucial feature together with the decreasing cost of sequencing has allowed cfDNA to become one of, if not the most promising liquid biopsy technique.

As its name suggests, cfDNA is extracellular DNA present in biofluids, particularly blood, that originates mainly from dying cells as part of normal cellular turnover or pathological processes, including cancer [84]. The presence of circulating tumour DNA (ctDNA) was first proposed in the late 1980s [85, 86], and formally proven in 1996 by Nawroz et al. [87], who found matched mutations between cfDNA and primary tumour. Today, over 1000 clinical trials are exploring the potential of cfDNA in oncology [88]. While most of these studies are focused on a specific cancer type, cfDNA can also be used for multi-cancer early detection, as prototyped in the Circulating Cell-free Genome Atlas study [89]. The clinical utility of the test is being evaluated using observational cohort [90] and prospective randomized controlled trial [91].

1.3 Biology of DNA methylation in human

There are many types of DNA modifications in nature, but DNA methylation in human genetics usually refers specifically to 5-methylcytosine (5mC). 5mC is a very special DNA modification in that it can be stably transmitted across cell division, or even across generation, via DNA methyltransferase 1 (DNMT1) in mammalian cells [92], which copies the epigenetic mark of the template strand onto the newly synthesized daughter strand during DNA replication at the exact same nucleotides. In this section, I will review the cellular mechanisms involved in DNA methylation and the functional implications with human relevance.

1.3.1 Maintenance of DNA methylation

The mechanism for the highly precise maintenance of 5mC is, in my honest opinion, one of the most elegant things in molecular biology. The trick is that inheritable 5mC occurs in a special dinucleotide context called "CpG", which stands for cytosine (C) and guanine (G) with a phosphate backbone (p) in between. I cannot find the exact source of this abbreviation, but I believe it was first used by Smith and Markham in their investigation of phosphate ester links in DNA [93, 94], and researchers kept the CpG notation to avoid confusion with CG base pairing. Because of this pairing, CpG dinucleotide is palindromic, meaning that the opposite dinucleotide is also CpG when read from 5' to 3'. Normally, both CpGs are symmetrically methylated. During replication, an unmodified C is incorporated, forming a hemi-methylated CpG, which is recognized by DNMT1 for further methylation.

There are several important functional domains in DNMT1, namely the target recognition domain (TRD), catalytic domain, CXXC domain, and replication foci-targeting sequence (RFTS) domain. DNMT1's specificity for hemi-methylated CpG comes from TRD, which is a subdomain of the active catalytic domain. Upon binding, the catalytic domain methylates the newly incorporated C. The recognition specificity of the TRD is great but not super, and demonstrates some degree of *de novo* methylation activity in unmethylated CpG under *in vitro* condition [95],

which can be detrimental if happened *in vivo*. Fortunately, additional mechanisms exist in DNMT1 to modulate its specificity.

One mechanism comes from CXXC domain, which has high affinity for unmethylated CpG. When DNMT1 binds to unmethylated CpG, a conformational change of the protein occurs and blocks the DNA from entering the catalytic pocket [96]. This intrinsic component ensures that DNMT1 possesses no *de novo* activity.

Another repressive mechanism comes from RFTS domain and its cooperation with UHRF1, and has to do with the coupling of DNMT1 with DNA replication. It was proposed that DNMT1 has an active and an inactive state. In the absence of DNA substrate, crystal structure of DNMT1 showed that RFTS is inserted deeply in the catalytic pocket [97], and no methyltransferase activity could happen. UHRF1, on the other hand, is a multi-functional protein with E3 ligase activity [98]. It is required for the inheritance of DNA methylation as shown in knockout experiments in mouse embryonic stem cells, and like DNMT1, also possesses the ability to bind hemi-methylated CpG via its SRA domain [99, 100]. UHRF1's expression is tightly coupled with the cell cycle [98] and has an affinity for trimethylation at lysine 9 of histone H3 (H3K9me3) through its TTD domain, but do not require H3K9me3 to bind chromatin [101]. It deposits monoubiquitin at lysine K14, K18, and K23 on histone H3 [102–104]. The multi-monoubiquitination then acts as a docking site for the RFTS domain of DNMT1, pulls RFTS out of the catalytic domain, and turns DNMT1 into its active form [102]. Once activated, DNMT1 operates in a processive manner, and scans for hemi-methylated CpGs along the DNA [105]. When RFTS is deleted, DNMT1 activity is no longer UHRF1-dependent and may lead to aberrant methylation [106], although the exact biological relevance have not been investigated to the best of my knowledge.

1.3.2 Dynamic DNA methylation changes

DNA methylation is not entirely about copying existing marks, otherwise it would be boring. What is truly amazing about DNA methylation is that certain loci can respond to environmental changes, and pass it down to the next generation [107],

somewhat resembling Lamarck's theory of evolution. DNA methylation also differs between different cells of the same individual. To be able to achieve this, there must be mechanisms that allow adding *de novo* methylation or removing existing ones.

De novo methylation

The *de novo* CpG methyltransferases DNMT3A/B in mammals were first identified in 1998 [108], and unlike DNMT1, they showed equal preference to unmethylated and hemi-methylated DNA, which can be explained by the structure of *de novo* methyltransferase, as the non-target C is not in contact with the enzyme [109]. The catalytically inactive DNMT3L is also required as a cofactor for the methyltransferase activity in germ cells [110], or DNMT3B3 for the case in somatic tissues [111, 112].

Similar to DNMT1, DNMT3A/B also have a regulatory mechanism via its ADD domain, which binds to the catalytic domain thus preventing enzymatic activity [113]. In the presence of H3, the ADD domain would bind to the unmethylated K4, allowing DNA to enter the catalytic domain. However, if K4 is methylated (H3K4me), ADD remains bound to the catalytic domain, and DNMT3A/B remains inactive. As H3K4me is generally associated with active transcription, this mechanism can protect active genes from being methylated.

Several differences exist between DNMT3A and DNMT3B, despite having high sequence similarity, suggesting different biological functions. As a supporting evidence, DNMT3A [114] and DNMT3B [115] have different disease manifestations when mutated in human. The first major difference between the two enzymes is that, although both have strong preference for CpG, they have differential preference for the flanking +1 base, with DNMT3A preferring C while DNMT3B preferring G/A [116, 117]. Another major difference is the mode of enzyme activity. DNMT3B, like DNMT1, is a processive enzyme and diffuse linearly along the DNA molecule [118, 119]. On the other hand, DNMT3A is a distributive enzyme [118, 120], meaning it binds CpG stochastically. Yet, due to its unique biochemical property, DNMT3A can form long stretches of DNMT3A-DNMT3L oligomers that binds to DNA,

resulting in nucleoprotein filaments detectable under scanning force microscopy [121–123], enhancing its methyltransferase activity exponentially [118].

DNMT3A/B also possess a PWWP domain, which recognizes H3K36me and is involved in the recruitment of these methyltransferases at gene bodies. This process will be further elaborated in later sections.

Active demethylation

Direct removal of methyl group from 5mC is a thermodynamically unfavourable reaction, and so far no convincing evidence exists to support such reaction [124]. The current understanding of active removal of 5mC involves iterative oxidation from 5mC to 5hmC (5-hydroxymethylcytosine), 5fC (5-formylcytosine), and 5caC (5-carboxycytosine) by ten-eleven translocation (TET) enzymes [125]. The oxidative derivatives 5fC and 5caC can be converted to abasic sites by thymine-DNA glycosylase (TDG), which are subsequently repaired via the base excision repair (BER) pathway to incorporate unmodified C [126, 127]. Alternatively, because DNMT1 binds poorly to hemi-5hmC [128], the opposite CpG to 5hmC would not be methylated, and DNA methylation would be gradually lost upon repeated cell division, avoiding the need to further oxidize 5hmC, which is not an efficient reaction [129]. The most notable mammalian examples of active demethylation are the two waves of global demethylation in pre-implantation embryo and gametogenesis, both mediated by TET enzymes [130].

There are 3 TET enzymes, namely TET1, TET2, and TET3. TET1 and TET3 possess a class of CXXC domain which recognizes CpG, but does not specifically bind unmethylated CpG, and may have affinity towards CpH sites as well (H denotes non-G bases) [131, 132]. On the other hand, the CXXC domain of TET2 is separated from its ancestral version likely via a chromosomal inversion during evolution, and formed CXXC4 (also named IDAX), a gene 650 kb upstream to the TET2 gene, and is implicated in the antagonistic regulation of TET2 [133]. The same paper also showed that the CXXC domain of TET3 may be regulating TET3

similarly [133]. Therefore, akin to the case of DNMT1, CXXC domains of TET enzymes seem to play a regulatory role rather than target recognition.

For target recognition, Hu et al. showed that the catalytic domain of TET interacts only with CpG of the target strand, and the complementary bases are not in direct contact with TET [134], suggesting that the methylation status of the non-target strand may have little effect on TET's binding. The methyl group of the target 5mC also do not interact with any protein residues (though it does interact with the catalytic Fe(II) ion) [134], which the authors proposed to be the reason why TET can accommodate different 5mC derivatives as substrates for further oxidation. Indeed, the same group solved the crystal structures of TET2 with 5hmC/5fC and demonstrated a near-identical conformation within the catalytic site [129].

Using BER as a means of demethylation is rather risky as it involves creating single strand breaks. Especially given that CpG methylation is symmetrical, there is a chance of creating double strand breaks (DSBs) which are highly deleterious. To avoid this, the cell must have a mechanism to complete the demethylation of one 5mC before moving on to the next. Recent studies by Weber et al. [135] and Liu et al. [136] may shed light on this subject. Previous work have shown that UHRF2 is a specific 5hmC reader [137, 138] and have ubiquitin ligase activity, which can be activated by hemi-hydroxymethylated CpG but not symmetrically hydroxymethylated CpG [139]. Weber et al. showed biochemically that TET interacts with TDG physically to convert 5mC to abasic site, and TDG can cooperate with the BER system to efficiently replace 5caC with C with minimal DSB products. However, they failed to demonstrate directly whether TET-TDG-BER can convert symmetrically methylated CpG into C with minimal DSBs. How TDG cooperate with BER was also unknown. On the other hand, Liu et al. showed that activated UHRF2 can catalyse non-proteolytic polyubiquitination of XRCC1, a core member of the BER machinery, and recruit TDG via an adapter protein called RAD23B. This links with the work by Weber et al., and suggests a multistep model where TET first performs an initial oxidation to form hemi-5hmC, which activates UHRF2, followed by polyubiquitination-mediated recruitment of the TET-TDG-BER complex to

further oxidize and excise the 5hmC, and ultimately result in the asymmetric demethylation of 5mC without DSB formation. Perhaps intriguing though, the 5hmC-containing oligonucleotides used for *in vitro* activation of UHRF2 in this study were not in CpG context.

Finally, like UHRF1, UHRF2 has functional domains that are known to interact with histone modifications, in particular H3K9me [140]. Therefore, the work by Liu et al. also provides a possible mechanistic link between histone modifications and active demethylation of 5mC. DPPA3 (also known as PGC7 or Stella) has been reported to protect the maternal genome from active demethylation by binding to H3K9me post-fertilization [141], as well as to protect from active methylation by excluding UHRF1 pre-fertilization [142]. Given the structural similarity between UHRF1 and UHRF2, it may be interesting to investigate whether DPPA3 prevents active demethylation by excluding UHRF2.

1.3.3 Functional organization of DNA methylation

Throughout the years our toolbox to quantitatively measure DNA methylation has evolved tremendously, from paper chromatography [143] to mass spectrometry [144], to the restriction enzymes HpaII and MspI [145, 146], to bisulfite conversion [147], bisulfite-free conversion [148–150], immunoprecipitation of 5mC [151], and direct measurement of 5mC by sequencing [152, 153]. Different method measures DNA methylation on a different scale and resolution, which is accompanied by the changing paradigm over the years of how researchers think about DNA methylation. Even today we are still learning about the roles of DNA methylation, what are its meanings in different cellular contexts, and more fundamentally, what are its functions in different genomic contexts.

Gene silencing by promoter methylation (less than 5 kb)

The first identified function of 5mC in eukaryotes is perhaps gene repression via promoter methylation. Notable examples in cancer are the methylation silencing

of tumour suppressor genes, such as CDKN2A [154], VHL [155], MLH1 [156], and BRCA1 [157].

Back in 1980, Liskay and Evans showed that cell transformation using HPRT gene from the inactive X chromosome cannot be expressed, whereas that from the active X chromosome can [158]. Since purified DNA was used, the authors concluded that the differences between active and inactive X chromosome must lie in the DNA itself, and proposed either DNA methylation or base substitution was responsible. Soon, Fradin et al. compared the production of viral proteins after injecting artificially methylated versus unmethylated viral DNA into frog oocytes [159]. Only the viral gene with 5' methylation was specifically repressed, whereas other genes were unaffected. Subsequent studies by Watt et al. further showed that only 5' methylation at specific CpG sites can suppress gene transcription, but not methylation elsewhere [160]. Due to the sequence specificity, they proposed that DNA methylation may directly block transcription factor binding, and supported their hypothesis by showing lack of binding activity of the respective transcription factor to methylated DNA [160]. This theory is later on shown to be insufficient, as not all transcription factors possess CpG at their recognition sites. Rather, the inhibitory effect by 5mC is more generally mediated by methyl-CpG binding proteins which repress transcription [161–163]. Although these early experiments were done using viral promoters and do not fully recapitulate endogenous promoter elements, they provided the much-needed mechanistic insight to understand the role of DNA methylation in eukaryotes. Indeed, an endogenous exemplar was soon discovered by Yen et al., who showed that the 5' region of HPRT were differentially methylated between active and inactive X chromosomes [164].

In 1985, Bird et al. described the existence of dense clusters of unmethylated CpG [165], later known as CpG islands (CGIs). The discovery of CGI is unusual because the majority of CpG is methylated [94], and the frequency of CpG in the genome is generally low [166] due to the mutability of 5mC into T [167]. Bird et al. further calculated the ratio of CpG compared to the local GC content, and found that unlike the remaining genome that has a depletion of CpG given the GC content, there is

no depletion nor enrichment of CpG in CGI. This led to the authors' conclusion that these GC-rich regions, rather than CpG itself, are somehow protected from methylation in the germline throughout evolution, and hence the unmethylated CpG were spared, leading to a relative enrichment within CGI compared to the rest of the genome. Bird also observed that quite a few housekeeping genes were associated with CGI, and found evidence that artificial manipulation of methylation in these CGI can modulate gene expression [168]. It was therefore suggested that CGI may be considered as a functional element that allows nearby genes to be expressed. This paradigm is still widely recognized nowadays, though many exceptions exist. The mechanism that guards CGI from methylation also remains elusive.

Methylation in complex regulatory networks (less than 5 kb)

The working model of gene silencing by promoter methylation is easy to comprehend. However, promoter is not the only regulatory element, and often times more complex mechanisms are in play. Here, I will highlight a few notable examples implicated in human disease.

Controlling chromatin boundary. The case of H19/IGF2 is famous for being controlled by genomic imprinting. Curiously, the epigenetic regulation of the two genes is controlled by the same region known as the imprinting control region (ICR) that lies between the two genes, which when methylated silences H19, and when unmethylated silences IGF2. Hark et al. [169] carefully dissected the ICR and revealed that this is due to the methylation-sensitive binding of CTCF, a zinc finger protein that is now known to play a key role in defining chromatin boundaries by mediating the formation of chromatin loops [170]. Briefly, IGF2 and H19 expression are both dependent on enhancer activation. When the ICR is unmethylated, the chromatin containing IGF2 is "insulated" from that containing H19 and an enhancer, and only H19 has access to the enhancer, thus activating H19 but suppressing IGF2. The ICR also dual-function as a 5' promoter-like element for H19. When the ICR is methylated, H19 is directly suppressed whereas IGF2 gains access to the enhancer, thus achieving the reciprocal gene regulation [171].

Double negation. UBE3A is another imprinted gene with maternal allele expression in neuronal cells, and methylation of the control region activates gene expression through a double-negation mechanism. Instead of directly modulating UBE3A, DNA methylation regulates the expression of an antisense long non-coding RNA, SNHG14 (also known as UBE3A-ATS) [172], which extends into the UBE3A gene region. This interferes with UBE3A transcription in *cis*, with only the 5' end being transcribed [173]. Several models have been proposed to explain how the antisense RNA truncates the UBE3A transcript, with the mainstream hypothesis being transcriptional collision [172]. However, antisense oligos have been shown to reverse UBE3A inhibition [174], which suggests that the effector may be the antisense RNA itself rather than the transcriptional machinery. Thus, I believe that alternative hypotheses such as R-loop formation or certain RNA binding proteins may be more appropriate [175].

Suppression of alternative promoters. It was already noted in the 1990s that while promoter methylation is negatively associated with transcription, gene body methylation is positively associated with transcription [176]. It turns out that in the latter case, methylation is a consequence of active transcription by *de novo* DNA methylation [177, 178]. During transcription, RNA Pol II recruits histone methyltransferases including SETD2 and deposits H3K36me [177, 179], which can be recognized by the PWWP domains of DNMT3 and thus result in *de novo* DNA methylation [177, 180]. Without the transcription-coupled deposition of DNA methylation, alternative transcription may occur by the engagement of RNA Pol II to intronic promoters [178]. This therefore highlights a potential role of DNA methylation in regulating transcript variants. One example in cancer is the LEF/TCF family genes; the full length transcripts contain a β-catenin binding domain and can drive the Wnt signalling pathway, whereas the truncated transcripts using intronic promoters do not, resulting in opposite actions [181]. It has been reported in colon cancer that there is aberrant activation of the full length transcript accompanied by the silencing of the short variant transcript [182],

leading to the misregulation of the Wnt pathway. Albeit DNA methylation was not measured in the study, it is very possible that the intronic promoter is silenced via the mechanism discussed above.

Transcriptional read-through and spillover methylation. The last case I would like to discuss is a special heritable form of MSH2 silencing, and may be related to the mechanism of transcription-coupled DNA methylation. MSH2 is a mismatch repair gene, mutation or silencing of which results in Lynch syndrome, a genetic disorder with increased risk of multiple cancer types. It has been reported that deletion of the last exon of EPCAM, a cell adhesion molecule, leads to MSH2 promoter methylation and silencing [183]. It appeared that MSH2 is located around 15 kb away from the 3' end of EPCAM, and the partial deletion in EPCAM resulted in a loss of polyadenylation signals and therefore transcriptional read-through into MSH2 gene. The relationship between transcription and DNA methylation was not clear at the time, but the authors noted that a similar scenario of read-through transcription and methylation silencing that was reported before in α-thalassemia [184], which strengthened their belief on the causal relationship. Now with the mechanistic insight of gene body methylation revealed, it could be interesting to revisit hereditary diseases associated with 3' end deletion of genes.

Methylation canyon or valley (more than 5 kb)

With the advances in sequencing technologies and growing interests, researchers began to investigate methylation in a relatively broader landscape. It was first reported by Long et al. the existence of large, evolutionarily conserved organization of lowly methylated regions that often covers the entire gene [185]. It was also discovered that genes associated with these large regions of low methylation are enriched in developmental genes, which are therefore associated with H3K27me3 marks [185, 186]. Xie et al. independently described these regions in the same year from a developmental perspective and coined the term "DNA methylation valley" (DMV) [187]. Later on, Jeong et al. also described similar regions as "DNA methylation canyon" in haematopoietic stem cells [188]. Although being

unmethylated, genes associated with DMVs are not necessarily active, but rather regulated by histone-mediated epigenetic mechanisms [188] such as silencing by polycomb repressive complex (PRC). Interestingly, it has been reported that prolonged *in vitro* culture can lead to hypermethylation of developmental genes [189]. Cancers also frequently gain DNA methylation in PRC-silenced developmental genes [190, 191], which may account for the CpG island methylator phenotype (CIMP) [192]. CIMP was first described in colorectal cancer and is thought to be one of the carcinogenic pathways in sporadic colorectal cancers by silencing of mismatch repair genes [193]. However, I manually checked the methylation status of mismatch repair genes (MSH2, MSH3, MSH6, MLH1, MLH3, and PMS2) and none of them associates with DMV nor H3K27me3 in normal tissues (data not shown). It is therefore demonstrated that CIMP is likely more than just hypermethylation of polycomb targets. Details of the methylation characteristics in cancers will be covered in a later section.

Partially methylated domain (more than 10 kb)

In my review of the historical literature on DNA methylation, there seemed to be a general notion that DNA methylation is associated with heterochromatin. This notion perhaps came from the fact that methylation is usually a repressive mark, which should be more abundant in the inactive chromatin. Comparison of 5mC distribution with karyotype images showed co-localization with constitutive heterochromatin [194]. MeCP2, a methyl-CpG binding protein, also co-localize with constitutive heterochromatin [195]. Early studies on X inactivation showed the key role of DNA methylation in maintaining the silencing [196]. UHRF1, a protein essential for the maintenance of DNA methylation, specifically recognizes H3K9me3 [197], which is a key histone modification in forming heterochromatin. Given all these examples, it is understandable why heterochromatin is thought to be enriched in DNA methylation.

Therefore, the work by Weber et al. showing that the inactive X chromosome is overall hypomethylated compared to its active counterpart is rather surprising [151].

Unlike previous studies that are limited to low resolution technique or CpG islands, Weber et al. developed methylation analysis by DNA immunoprecipitation (MeDIP), which allowed a relatively unbiased assessment of genome-wide methylation. This finding was later confirmed by Hellman and Chess [198], who showed a higher level of methylation in the active X, which is not present prior to X inactivation in embryonic stem cells.

In 2009, Lister et al. published the first set of base-resolution human DNA methylome [199]. They described broad regions of reduced methylation level ($< 70\%$) in immortalized fibroblasts, for which they named partially methylated domains (PMDs) because of the intermediate methylation level. The authors found that PMDs were not limited to X chromosome, but were instead found widespread in autosomes as well, and were associated with a decreased gene expression [199]. The authors at that time had no clue of the mechanism behind the formation of PMD, which took almost a decade before it was finally figured out.

The mystery of PMD is actually hidden in a paper published in 1971 by Adams before even knowing the function of DNA methylation, who showed in mouse fibroblasts that early replicating DNA is methylated quickly, whereas late replicating DNA lags behind [200]. In the same study, Adams showed that late replicating DNA contains less methylation than early replicating DNA, which fits well with the description of PMD described by Lister et al. In 2005, Weber et al. reported that global hypomethylation in cancers, which was known since 1983 [201, 202], may be non-random and enriched in gene-poor areas. Between 2009 and 2015, Feinberg's and Irizarry's labs published a series of paper characterizing that tumour hypomethylation occurs in broad domains which are normally associated with repressive histone marks such as H3K9me [203–205]. Perhaps most interestingly, they showed that the hypomethylated domains are universal features across common solid tumuors, and may be present in premalignant tissues as well. It was also shown that ageing and sun-exposed skin samples have a similar hypomethylation pattern in inactive chromatin [206]. In 2012, Berman et al. made the connection between these tumour-hypomethylated domains and fibroblast-PMD, and found considerable

overlap between the two [207]. Because of the strong overlap, the authors kept the PMD terminology even though the methylation level is way lower than what was initially reported by Lister et al. A breakthrough was made in 2013 by Cruickshanks et al. who highlighted the association between tumour-hypomethylated domains and replication timing [208]. They showed that senescent cells have a strikingly similar DNA methylation landscape to cancer cells, with hypomethylation happening at late replicating domains. Cruickshanks et al. attributed this to the failure of DNMT1 to maintain DNA methylation, and supported their hypothesis by demonstrating a reduced level of DNMT1 transcript and abnormal subnuclear localization of DNMT1 protein. The story finally came to an end in 2018, when Zhou et al. linked together PMD and late replication in mice and human *in vivo* data [209]. Zhou et al. demonstrated a progressive emergence of PMD across tissues sampled at different developmental stages, as well as a linear association between methylation levels in PMD and chronological age. In cancer, PMD hypomethylation is associated with tumour mutational burden, which the authors used as a proxy for mitotic clock. With strong *in vitro* and *in vivo* evidence, it is now recognized that the level of DNA methylation in repressive heterochromatin is generally reflective of the cell's mitotic history, and PMD hypomethylation is responsible for what was known as global hypomethylation since the 1980s.

However, there are some questions that remain unanswered.

1. If the historic literature by Adams [200] is correct, what causes the slower rate of replenishing DNA methylation in late replicating domains?
2. Cruickshanks et al. used senescent cell culture to model cancer methylation landscape, and showed aberrant DNMT1 behaviour [208]. However, the mechanism behind aberrant DNMT1 behaviour was not explored in detail. They couldn't detect DNMT1 on Western blot, but were able to detect DNMT1 in immunofluorescence imaging. Two anti-DNMT1 antibodies were used, one targets the N terminal of DNMT1 (ab19905), whereas the other targets somewhere around the middle of the protein (A300-042A). Could a truncated DNMT1 isoform either lacking the RFTS domain or lacking the

catalytic domain be expressed in senescent cells, leading to the failure of methylation maintenance? If so, do cancer cells behave the same?

3. Finally, now that the mechanism behind global hypomethylation in tumour has been worked out, how about the mechanism for focal hypermethylation?

The answer to (1) may be revealed by work from Nishiyama et al. [210], who showed that the maintenance of DNA methylation during early and late replication utilizes two different protein systems, which could explain the difference in rates of replenishing 5mC marks. Hypothesis (2) is certainly a very interesting one, but is beyond the scope of the current thesis. I aim to explore questions (3) via computational means.

1.4 Molecular characteristics of OAC

Given the heterogeneous nature of cancers, their molecular characteristics have been one of the central matters of both scientific curiosity and clinical utility, potentially informative of the carcinogenic pathways as well as treatment strategies. Long term and international efforts were spent on gaining pan-cancer and cancer-specific insights [211, 212], and OAC is of course not an exception. In this last review section, I will give a brief overview of what has been known about the key molecular characteristics of OAC, and is by no means comprehensive.

1.4.1 Aneuploidy

It has been long known that aneuploidy is common among BO with dysplasia and OAC, and may be a distinguishing feature between BO with low versus high risk of progression [213]. In general, copy number changes increase as disease progresses [214]. Subsequent retrospective and longitudinal cohorts demonstrated the potential clinical utility of using aneuploidy as an ancillary biomarker for BO progression [75, 215, 216]. It has also been shown that whole genome duplication [217] and genomic catastrophe events such as chromothripsis [218] are very common in OAC (around 60% and 30% respectively). Extrachromosomal DNA carrying oncogenic drivers is

also common in BO and OAC, and similar to chromosomal DNA, increases in copy number as the disease progresses [219]. The mechanism of increased copy number may be due to random segregation during cell division and subsequent positive selection due to proliferative advantage from oncogenes [220].

Reported recurrent loss of heterozygosity (LOH) events include TP53, CDKN2A/2B, SMAD4, SMYD3, ARID1A, ATM, APC, PTEN, and RUNX1 [218, 221–224]. Some very long genes (> 1 Mb) in fragile sites such as FHIT and WWOX are also recurrently affected, but the functional relevance is uncertain. In particular, LOH in TP53 and CDKN2A are suggested to be early events and fuel further chromosomal instability [217, 225, 226]. The landscape of recurrent amplifications includes MYC, KRAS, EGFR, ERBB2, GATA4, GATA6, CCND1, CDK6, etc.

1.4.2 Mutations

OAC has relatively high mutation burden compared to other cancers, but the prevalence of recurrent point mutations apart from TP53 and perhaps CDKN2A remains low. [223, 227]. However, it has been observed that certain mutations often coexist, whereas some are often mutually exclusive [223], potentially representing different carcinogenic pathways.

Perhaps more interesting is mutational signature analyses, which focuses on the underlying mutagenic aetiology rather than mutation of specific genes. Each mutational process leaves a characteristic pattern, the most well known example being ultraviolet radiation resulting in CG to TA mutations [228]. By looking beyond driver genes and assessing the mutational pattern genome-wide, we may detect the cancer's exposure to different mutagenic processes.

BO and OAC are enriched in T>C or T>G mutations at CTT context [224, 229, 230], known as SBS17a and SBS17b respectively, though the underlying mutational mechanism has not been delineated yet. Other notable mutational signatures include SBS1/5 (ageing related), SBS2 (APOBEC activity), SBS30 (BER related), SBS3 (homologous recombination related), and SBS44 (MMR deficiency). SBS44

is particularly interesting in the context of this thesis, because MMR deficiency tumours respond exceptionally well to immunotherapy.

1.4.3 DNA methylation

As mentioned in introduction, it has been known for very long that DNA methylation changes are early events in the development of BO and OAC. However, early studies focused on a targeted gene panel, and it wasn't until the 2010s when large-scale whole-genome methylation profiling was being performed using methylation arrays [5, 8, 222, 231–235] or bisulfite sequencing [236]. The analysis of genome-wide DNA methylation in cancer is largely influenced by the observations in colorectal cancer, where tumours were defined into CIMP subtypes based on the degree of hypermethylation of CpG sites that are ubiquitously unmethylated in normal tissues. Likewise, several papers have classified OAC by the degree of CpG hypermethylation [5, 231, 234, 235]. Some performed integrated analyses with other gastrointestinal track cancers and found OAC to be similar to a subtype of gastric cancer [222, 235]. Finally, recent studies included a wider selection of CpG sites, and did not limit their analyses to just unmethylated CpG islands in normal tissues. One group classified BO and OAC using non-negative matrix factorization (NMF) into 4 subtypes, and for the first time highlighted a subtype of OAC with predominantly hypomethylation [8]. Another group studied OAC and OSCC in the context of broad domains such as PMD and HMD (highly methylated domains), and found distinct methylation patterns between OAC and OSCC [236].

However, as briefly mentioned in the beginning of this chapter, all of these studies on cancer DNA methylation is based on bulk assays without considering the cellular contribution of the epigenetic mark. This poses a great challenge in making proper inferences from the data, and in my honest opinion a lot of the classification effort may be misleading. A dissection and demonstration of the issue will be presented in detail in Chapter 3, which forms an important aspect of this thesis.

2

Data generation and description

Contents

2.1 LUD2015-005 Trial design

The LUD2015-005 trial is an open-label phase I/II study to evaluate the safety and efficacy of combined ICI+CTX in lower oesophageal cancer [24]. The trial consists of 4 cohorts, and in this thesis I focused on the analysis of cohorts A/B, which are patients with metastatic or inoperable OAC. The analysis of scRNA-seq

data of this cohort focusing on the tumour microenvironment has been published recently [237] and described the trial design.

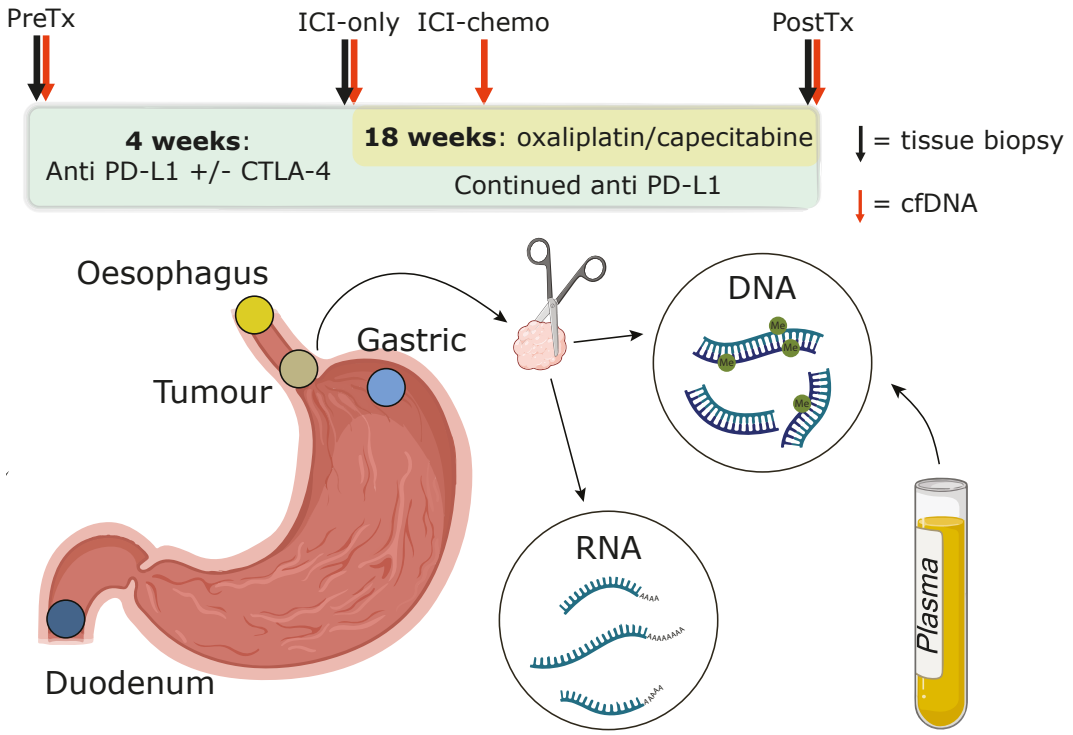


Fig. 2.1: Schematic diagram describing the LUD2015-005 trial design and sample collection timepoints (top). Biopsy samples were processed for DNA and RNA sequencing, and plasma was processed for cfDNA sequencing.

Briefly, the inoperable cohort has 3 treatment arms based on a dose-escalation strategy to assess the safety profile of combined durvalumab (anti-PD-L1) and tremelimumab (anti-CTLA-4). All patients received a fixed component of biweekly intravenous durvalumab (750 mg). Patients in the first treatment arm receives no tremelimumab. Those in the second arm received durvalumab plus a single 37.5 mg priming dose of tremelimumab. Patients in the third arm received durvalumab plus 75 mg tremelimumab. After 4 weeks of ICI treatment, patients received a maximum of six cycles of chemotherapy (CTX) with oxaliplatin and capecitabine, in addition to continued biweekly durvalumab until the end of treatment.

2.2 Outcome assessment

The primary outcomes of this trial are safety related. However, this thesis is more interested in the secondary outcomes related to treatment efficacy. These include overall survival (OS), progression-free survival (PFS), and objective tumour response measured using irRECIST criteria [238] (not to be confused with iRECIST [239], which was established after the trial has already started). Baseline computerized tomography (CT) imaging was conducted prior to treatment initiation, and every 6 week thereafter during treatment for the radiological assessment of tumour response. The irRECIST requires the definition of a set of "target lesions" at baseline, which would be used in subsequent assessments. Response categories are complete response (irCR, complete disappearance of lesions), partial response (irPR, $> 30\%$ reduction of measured target lesions in absence of progressing new or non-target lesions), progressive disease (irPD, $> 20\%$ increase in measured target lesions from baseline or nadir, or confirmed progression of new or non-target lesions), and stable disease (irSD), an intermediate category that captures the remaining cases.

A *post hoc* exploratory outcome binary metric was also defined by a centralized review process, termed clinical benefit (CB). Patients with ≥ 12 months PFS were deemed to have CB (n = 9), whereas those who progressed within 12 months (n = 13) were deemed to have no clinical benefit (NCB). One patient received alternative therapy within 12 months of treatment onset, and was excluded from analyses involving CB outcome.

2.3 Sample collection timepoints

There are 5 timepoints where sample collection may happen. The first timepoint is pre-treatment baseline (PreTx), the second is upon completion of 4 weeks of ICI (ICI-only), the third is at week 3 of the first cycle of combined ICI+CTX, the forth is upon completion of ICI+CTX (PostTx), and the last timepoint is the recall of long-term survivors (LTSR), who had exception responses to the treatment and kindly donated blood for the study after completion of the trial.

2.3.1 Tissue biopsy

Endoscopic biopsy samples were taken via oesophagogastroduodenoscopy (OGD) at PreTx, ICI-only, and PostTx. However, only results from the PreTx biopsies were used for analysis in this thesis.

At each OGD, up to five 2 mm biopsy pairs were taken from the site of the tumour, while 1-2 pairs of normal control tissue were also taken from normal oesophagus (at least 2 cm proximal to the GOJ or proximal extent of the cancer lesion, whichever is more proximal), gastric cardia (at least 2 cm distal to the GOJ or distal extent of the lesion), and the descending duodenum (D2). Each biopsy pair was split into two aliquots, one of which was immediately snap-frozen using dry ice or liquid nitrogen, while the other was slow frozen with 1 mL of fetal bovine serum (FBS) or heat-inactivated human serum with 10% DMSO and placed at $-80\,^{\circ}\mathrm{C}$ in a controlled-rate freezing container. Snap and slow-frozen biopsies collected at local enrolment sites were stored at $-80\,^{\circ}\mathrm{C}$, and were shipped on dry ice to the central site for long-term storage at $-80\,^{\circ}\mathrm{C}$ and subsequent processing.

2.3.2 Blood

Up to 20 ml of blood samples were taken in EDTA tubes at PreTx, ICI-only, ICI-chemo, PostTx, and LTSR. The blood was placed immediately on ice and transferred to the local laboratory, where tubes were centrifuged at 2000 g for 10 minutes at $4\,^{\circ}\mathrm{C}$. The plasma layer is aspirated without disturbing the buffycoat layer, and centrifuged again at 16000 g for 10 minutes at $4\,^{\circ}\mathrm{C}$. The plasma is then transferred to cryovials without disturbing the pellet and stored at $-80\,^{\circ}\mathrm{C}$, and were shipped to central site on dry ice if collected in local enrolment sites.

2.4 Sample processing

2.4.1 Bulk RNA-seq

The majority of bulk RNA-seq were performed by TM Carroll and JA Chadwick. Total RNA was extracted from snap-frozen biopsies using the mirVana miRNA

Isolation Kit (AM1560) per manufacturer's protocol, followed by incubation with TURBO DNase (AM1907) to remove genomic DNA contamination. RNA was then purified and concentrated using AMPure XP beads (A63880) at 2.8x ratio. RNA was quantified using Qubit RNA BR Assay and the RNA integrity number (RIN) was calculated using Agilent RNA 6000 Pico Kit (5067-1513). Bead-based rRNA depletion library preparation was performed with the TruSeq Stranded Total RNA Library Prep Human/Mouse/Rat Kit (20020596), and sequenced on Illumina HiSeq (75 bp paired end) to a target depth of 50 million reads per library.

2.4.2 Paired tumour and normal WGS

The biopsy samples were first embedded in Optimal Cutting Temperature compound (OCT) (361603E) for 10 to 20 cryosections at 10 µm. Remaining tissues were then melted in sterile phosphate buffered saline (PBS), followed by DNA extraction using GeneJET Genomic DNA Purification Kit (K0722). DNA was then purified and concentrated using AMPure XP beads (A63880) at 1.8x ratio. Genomic DNA quality was assessed using Nanodrop and gel electrophoresis, and quantified using Qubit dsDNA HS assay kit. Extracted DNA samples were then sent out for library preparation using a PCR-free protocol and sequencing using the Illumina NovaSeq6000 at 60x target coverage for tumour and 30x for paired normal. Duodenal biopsy was used as germline control whenever possible. In cases where the endoscope failed to reach the duodenum (e.g. due to tumour obstruction), buffycoat was used. In cases where buffycoat was unavailable, normal oesophageal biopsy was used.

Cryosections were performed by I Ratnayaka, tissue embedding and DNA extraction were performed by J Chang and me.

2.4.3 Tissue methylation sequencing

TET-assisted pyridine borane sequencing (TAPS) [148] was performed on all tumour biopsy samples and selected normal controls to obtain whole-genome DNA methylation at single-base resolution. Details of the protocol were previously described [148]. TAPS was performed by P Siejka-Zielińska, M Inoue, and me.

Briefly, genomic DNA was fragmented using Covaris M220 ultrasonicator, followed by size selection for fragments between 200-400 bp using AMPure XP. Adaptor ligation was performed using KAPA HyperPrep PCR-free Kit (KK8505) with methylated and unmethylated spike-in controls, followed by TAPS conversion using TET oxidation and borane reduction. The resulting libraries were then amplified for 4 cycles using KAPA HiFi HotStart Uracil+ReadyMix Kit (KK2802), and pooled together before sending out for sequencing using the Illumina NovaSeq6000 at 30x target coverage for tumour and 15x for normal tissues.

2.4.4 CfDNA methylation sequencing

Plasma samples were used for cfDNA extraction using QIAamp Circulating Nucleic Acid Kit, then quantified using Qubit dsDNA HS Assay. 1-3ml plasma was used to obtain $>$ 10 ng cfDNA per sample.

Similar procedures as tissue TAPS were performed for cfDNA. However, unlike tissue TAPS, due to the low amount of input DNA, the barcoded libraries were pooled together prior to TAPS conversion, and carrier DNA was used where library concentration was low to increase DNA recovery after conversion reaction. TAPS converted libraries were amplified for 4 cycles and sent out for sequencing using the Illumina NovaSeq6000 at 30x target coverage.

2.5 Bioinformatic pipelines

2.5.1 Bulk RNA-seq

Bioinformatic pipeline for RNA-seq was previously described [237]. Briefly, STAR [240] alignment was performed to a custom version of GRCh38 human genome and supplementary contigs, and GENCODE v34 GTF was used as a gene annotation file, filtering out readthrough transcripts and annotations to PAR regions in chromosome Y. The final counts matrix was produced using featureCounts [241].

Downstream differential expression was conducted using DESeq2 [242] using a significance threshold of $fdr < 0.05$ to identify differentially expressed genes unless otherwise specified. To account for the effect of varying tumour fraction (commonly

known as "purity") in the biopsy, purity was included as a covariate in all analyses performed. Details of the rationale will be explained in Chapter 3. The estimated tumour purity in RNA-seq samples were obtained from Carroll et al. [237] using BayesPrism [243]. Moderated gene-level log-fold changes were calculated within DESeq2 using the `ashr` method [244] using appropriate model contrasts. Gene set enrichment analysis was conducted using FGSEA [245] on MSigDB pathways [246]. Signed p values from DESeq2 were used to rank the genes. Log-normalized counts for visualization were generated using the `vst` function.

2.5.2 WGS

WGS sequencing reads were aligned to the hg38 human genome as previously described using BWA [247]. The `bwa mem` algorithm with ALT contig handling was used.

Short somatic variants were called using a consensus of at least two out of three callers (Strelka2 [248], Mutect2 [249] and Octopus [250]). Tumour mutational burden was calculated as the rate of non-synonymous mutations in coding region, which was 35.6 Mb in this study. Mutational signatures were evaluated using deconstructSigs [251] with the configuration described in [230].

Copy number alterations (CNAs) were called using Battenberg [252, 253]. Reproducible, semi-automated quality control of the CNA calls were performed using DPClust [253] and CNAqc [254] in a pipeline built by A Frangou and R Amess, published along with [237]. Ambiguities in CNA calls used in this thesis were resolved through manual inspection by myself with help from I Peneva. Whole genome duplication (WGD) was defined according to Dentro et al. [255], based on a cutoff on the ratio between estimated ploidy and percent genome with LOH. Genomic signatures of microsatellite instability (MSI) were detected using MSIsensor [256].

2.5.3 DNA methylation

TAPS sequencing reads were aligned to a customized hg38 human genome, created by concatenating the full hg38 reference with spike-in sequences. Alignment was done

using `bwa mem` because only 5mC was converted during the TAPS reaction, and the sequence complexity is adequate for four-letter aligners [148]. In-house benchmarking showed that using three-letter aligners designed for bisulfite sequencing such as HISAT-3N [257] resulted in same if not slightly worse alignment rate (performed by B Schuster-Böckler, data not shown). Methylation calling was performed using the `extract` subcommand from MethylDackel [258] for all CpG sites. Read-based methylation was called using `perRead` subcommand from MethylDackel.

Downstream differential methylation testing was performed using DSS [259] with considerable extension to account for the effect of CNA and tumour purity. Theory and details will be presented in Chapter 3, and therefore will not be described in full in this section. Analysis of read-based methylation will also be discussed in Chapter 3.

3

Computational methods and frameworks

Contents

3.1 Tumour fraction and copy number aware differential testing

3.1.1 Statistical framework

It is important to take tumour fraction into account when analysing bulk methylation sequencing data. Many methods to account for cell composition, copy number variation, or both have been developed for methylation array [260–262] and sequencing data, where the latter can be further classified into count-based approaches [263, 264] and read-based approaches [265–268]. However, despite the plethora of methods, only one tool [264] is both copy number and tumour fraction aware to the best of my knowledge. Furthermore, most of the packages are focused on cell-type deconvolution or calling tumour-specific methylation of individual samples, and do not consider the case for general experimental designs such as tumour-to-tumour comparisons. In this section, I will introduce the statistical frameworks I developed in analysing our dataset for this purpose.

Assume methylation is the same for all cells within each cell-type. Let there be m non-tumour cell types in a sample, with methylation $M_1, M_2 \dots M_m$ and proportion of DNA $\lambda_1, \lambda_2 \dots \lambda_m$. Let tumour methylation be M_t and DNA proportion be λ_t. Then, the observed overall methylation M_o can be expressed as:

$$\begin{aligned} M_o &= M_1\lambda_1 + M_2\lambda_2 + \cdots + M_m\lambda_m + M_t\lambda_t \\ &= \sum_{i=1}^{m} M_i\lambda_i + M_t\lambda_t \end{aligned} \tag{3.1}$$

In non-cycling, non-tumour cells, we can assume they are diploid, and therefore the corresponding DNA fraction λ is directly proportional to cell type abundance. If somehow the cell type proportions are known, and we have a reliable methylation atlas for reference M values, M_t can then be computed based on the linear additive property of DNA methylation shown in Eq. (3.1).

However, it should be noted that methylation atlases are usually constructed using tissues donated by healthy individuals [269]. In the context of tumour biopsy, the healthy reference may not be applicable to the complex tumour

microenvironment. More importantly, in most if not all cases, we do not have the ground truth of cell type proportions.

We can, however, estimate tumour fraction (commonly termed "cellularity" or "purity") using allele-specific copy number calling methods [252, 253, 270]. Briefly, copy number is always an integer, and based on the extent of allelic imbalance and change in sequencing coverage, these computational algorithms can find a best-fitting copy number profile and tumour cellularity.

Let ρ be the tumour cellularity, $1-\rho$ be the non-tumour cellularity, ψ be the local tumour ploidy for a certain genomic region, and 2 be the ploidy for non-tumour cells. By definition, the tumour DNA fraction λ_t equals the amount of tumour DNA divided by amount of total DNA. The expression is given by:

$$\lambda_t = \frac{\psi\rho}{\psi\rho + 2(1-\rho)} \tag{3.2}$$

To further simplify the equation, let M_n and λ_n be the average methylation and collective proportion of non-tumour DNA respectively. Since the proportion of tumour and non-tumour adds up to one, λ_n can be expressed as $1-\lambda_t$. Now that only the tumour DNA fraction λ_t variable remains, it will henceforth be referred to as λ. In other words, Eq. (3.1) can be rewritten as:

$$\begin{aligned} M_o &= M_n\lambda_n + M_t\lambda_t \\ &= M_n(1-\lambda_t) + M_t\lambda_t \\ &= M_n + (M_t - M_n)\lambda \end{aligned} \tag{3.3}$$

From Eq. (3.3), M_o can be expressed as a function of λ_t, which is a simple linear equation with slope $(M_t - M_n)$ and intercept M_n.

A widely accepted statistical model for finding differentially methylated region (DMR) between two comparison groups with sequencing data is beta-binomial regression. A computationally efficient algorithm which allows general experimental design has been implemented in the R package, *DSS* [259]. Unlike other methods

that use generalized linear models, *DSS* implements a special arcsine link function which allows the use of generalized least squares procedure.

By calculating λ from Eq. (3.2) for each copy number segment, we may fit a linear model according to Eq. (3.3) under the *DSS* framework, and obtain the estimated tumour and non-tumour methylation from the model coefficients. Subsequently, model contrast may be performed to obtain test statistics using Wald test on specified linear combinations of model coefficients.

Example 1: DMR in tumour versus non-tumour comparison. Consider sample set A, which has 15 samples with tumour DNA fraction λ_i uniformly distributed between 0 and 1. Let the ground truth tumour methylation $M_t = 0.1$, and non-tumour methylation $M_n = 0.8$, and an average sequencing depth of 30x. Detailed code for the simulation and DMR testing is available in Appendix A.1.

Using the *DSS* package, a model can be fitted using λ_i as a covariate (coded as `~ lambda` in R). In mathematical notation, the model design is:

$$M_o = \beta_0 + \beta_1 \lambda_i$$

Where β_0 and β_1 are the model coefficients to be estimated using beta-binomial regression. As mentioned, from Eq. (3.3) it can be seen that $\beta_0 = M_n$ and $\beta_1 = M_t - M_n$. The simulated data and fitted model are shown in Fig. 3.1.

Conceptually, DMR testing can be performed tumour methylation versus non-tumour methylation without the use of normal controls.

This can be done by testing the null hypothesis of $M_t = M_n$ using Wald test, which is equivalent to testing the model coefficient $\beta_1 = 0$:

$$\begin{aligned} H_0 : \quad & M_t = M_n \\ & M_t - M_n = 0 \\ H_0 : \quad & \beta_1 = 0 \end{aligned}$$

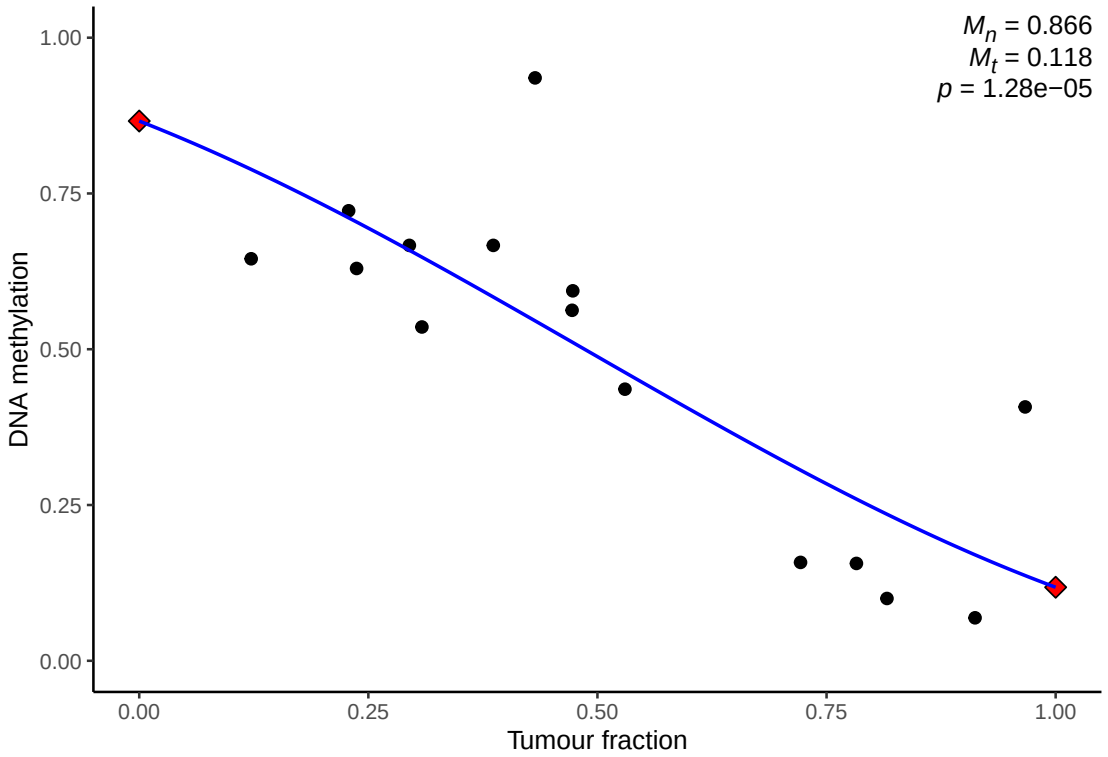


Fig. 3.1: Methylation is visualized as a function of tumour DNA fraction. Each black dot represents a sample, blue line represents the fitted beta-binomial model, and the diamonds at tumour fraction equals 0 and 1 represent the estimated non-tumour methylation M_n and tumour methylation M_t. Their values are 0.866 and 0.118 respectively. The p value is $1.28e{-}5$.

The estimated M_n and M_t are 0.866 and 0.118, both close to the ground truths of 0.8 and 0.1. In addition, $p < 0.05$ and therefore the null hypothesis is rejected, suggesting a significant difference between tumour and non-tumour methylation.

Example 2: DMR with additional covariates. Consider sample set B, which has a different ground truth tumour methylation $M_t = 0.8$, and non-tumour methylation $M_n = 0.6$. We may include subtype as an additional covariate in the model design, and analyse set A and B together.

Let dummy variable $S = 0$ for subtype A, and $S = 1$ for subtype B. Since DNA methylation of both the tumour and the microenvironment can be different between the two subtypes, we need to include interaction terms for both the slope

and intercept. Hence, the design formula is:

$$M_o = \beta_0 + \beta_1\lambda + \beta_2 S + \beta_3\lambda S$$

Or:

$$\begin{aligned} M_o &= \beta_0 + \beta_1\lambda && \text{for subtype A, where} \quad S{=}\,0 \\ M_o &= (\beta_0 + \beta_2) + (\beta_1 + \beta_3)\lambda && \text{for subtype B, where} \quad S{=}\,1 \end{aligned}$$

The possible DMR tests that can be performed include:

(1) Tumour subtype A versus tumour subtype B

(2) Non-tumour subtype A versus non-tumour subtype B

(3) Tumour subtype A versus non-tumour subtype A

(4) Tumour subtype B versus non-tumour subtype B

Comparisons (1) and (2) are shown in Fig. 3.2.

Let M_{tA}, M_{tB} be the tumour methylation of sets A, B and M_{nA}, M_{nB} be the corresponding non-tumour methylation. The null hypothesis for (1) is:

$$\begin{aligned} H_0 : \quad & M_{tA} = M_{tB} \\ & (M_{tA} - M_{nA}) + M_{nA} = (M_{tB} - M_{nB}) + M_{nB} \\ & \beta_1 + \beta_0 = (\beta_1 + \beta_3) + (\beta_0 + \beta_2) \\ H_0 : \quad & \beta_2 + \beta_3 = 0 \end{aligned}$$

And the null hypothesis for (2) is:

$$\begin{aligned} H_0 : \quad & M_{nA} = M_{nB} \\ & \beta_0 = \beta_0 + \beta_2 \\ H_0 : \quad & \beta_2 = 0 \end{aligned}$$

The estimated tumour and non-tumour methylation values for subtype A are 0.119 and 0.866 respectively. Note that the estimated values are slightly different

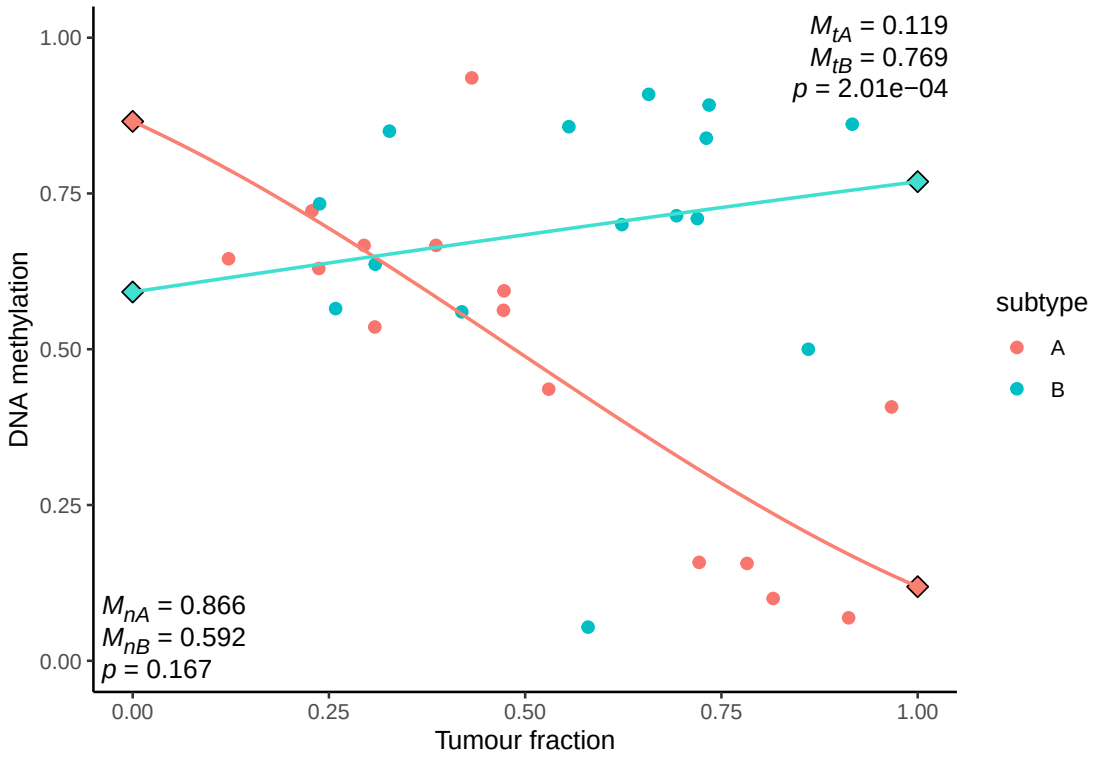


Fig. 3.2: Methylation is visualized as a function of tumour DNA fraction. Salmon colour represents set A and turquoise represents set B. Each dot represents a sample. Lines represent the fitted beta-binomial model. Diamonds at tumour fraction equals 0 and 1 represent the estimated non-tumour and tumour methylation.

from those in example 1, because sample subtypes A and B are modelled together, resulting in a slightly different dispersion estimate under the *DSS* framework. The corresponding values are 0.769 and 0.592 for subtype B, which are good estimates of the ground truth values of 0.8 and 0.6.

The p value for subtype A tumour versus subtype B tumour comparison is $2.01e{-}4$, suggesting a significant difference. The p value for subtype A non-tumour versus subtype B non-tumour is 0.167, suggesting that there is not enough power to reject the null hypothesis at the given sequencing coverage, sample size, and methylation difference.

The above illustrated differential testing framework can in theory be applied to all quantitative molecular signatures, including but not limited to transcriptomics, proteomics, and histone modifications, as long as the assumptions are fulfilled.

3.1.2 Genomic intervals for differential testing

The previous section proposed a framework for differential testing given some data with varying tumour DNA fraction. A structure for feature selection, i.e. defining the genomic regions for differential testing is also required. Unlike transcriptomics or proteomics which have universally-agreed annotations, epigenetic data often falls in non-coding regions where there are no defined boundaries.

A completely context-agnostic approach is to test each cytosine or CpG dinucleotides for the methylation level, assuming the methylation between each element is independent. Yet, it is well-known that the methylation of one CpG is highly correlated with its neighbouring CpGs. This can be explained by the oligomerization of DNMT3A or processivity of DNMT3B, both capable of *de novo* methylation over long stretches of DNA. Hence, it makes more biological sense to group CpGs in proximity together, which can increase statistical power and reduce the required sequencing depth.

The bioinformatics community has taken different approaches to this problem. *DNMTools* (formerly *MethPipe* [271] and *RADMeth* [272]) implements a stepwise process of segmenting the genome into hypomethylated regions, as well as testing each CpG genome wide, and finally using the hypomethylated intervals for combining the CpG-level differential testing. *BSmooth* [273] and *DSS* also test each CpG genome wide, but applies a smoothing function for a certain window around each tested site, followed by a post-hoc merging of significant CpGs into DMRs. *metilene* [274] uses circular binary segmentation with some heuristics to segment the genome into regions with similar methylation levels, followed by a two-dimensional Kolmogorov-Smirnov test. *metilene* is by far the most efficient DMR algorithm according to my personal experience, but its statistical design does not allow the inclusion of covariates. Finally, *wgbstools* [269], although not a DMR method, has implemented a segmentation algorithm for methylation sequencing data across multiple samples based on binomial likelihood maximization, which can then be used for downstream DMR testing using other methods.

The literature on DMR methods are ever-growing and there is no consensus to what is the best way of defining genomic intervals for *de novo* DMR discovery. Another approach is to use prior-defined genomic intervals that are known to be likely functional units, such as known promoters and enhancers. Most of these defined functional units are context-specific and do not cover the entire genome, thus wasting some information contained in the genome-wide dataset if used in DMR testing. However, a recent work has been published for a set of "full-stack" chromatin states across more than a hundred cell types using *ChromHMM* [275], which covers almost the entire hg38 human genome.

In this project, I defined three sets of genomic intervals to be tested for DMR:

(1) *wgbstools* segmentation using all available samples with programme defaults
(2) CGI as defined by the UCSC genome browser, and segmentation of the remaining non-CGI intervals into 2 kilo-base-pair (kb) windows with shifted boundaries where necessary such that no CpG dinucleotide overlap more than 1 window
(3) "Full-stack" chromatin states with boundary modifications due to reason mentioned in (2), as well as trimming overlapping genomic intervals due to artefacts created by lifting over from hg19 to hg38 reference genome

Unless otherwise mentioned, the DMR discussed in results would be those defined using chromatin states from (3), because it is a natural choice for the downstream annotation of DMRs.

3.1.3 Filtering of differential testing results

Outlier removal

A key assumption in group versus group comparisons is that each sample within the same group belong to the same distribution. For parametric statistical tests, normal distribution is assumed. In the *DESeq2* package [242] for differential gene expression testing, negative-binomial distribution is assumed. In the *DSS* package, beta-binomial distribution is assumed.

However, this assumption may not be true, and there may be outliers that heavily influence the regression model. To account for this, *DESeq2* implemented the Cook's distance diagnostic [276], which is calculated for each sample at each tested feature, and is a measure of how much the regression model changes if the i^{th}data is removed. A critical cutoff for Cook's distance at a desired type I error rate α can be calculated using an F-distribution with the appropriate degrees of freedom [242].

Inspired by *DESeq2*, I implemented the Cook's distance metric with great help from Dr. George Nicholson. Detailed derivation and equations can be found in Appendix B. Unless otherwise specified, $\alpha = 0.1$ is used, meaning that given there is no outlier, a genomic interval may still be rejected 10% of the time due to random chance. Genomic intervals with any comparison group with less than 3 samples excluding biological replicates are also removed. Then, p values are adjusted using the Benjamini-Hochberg method [277] for multiple comparisons within each model contrast across all genomic intervals. The adjusted p value will be referred to as False Discovery Rate (fdr) in this thesis. The critical threshold to declare significance is chosen to be 0.01.

Effect size threshold

Another commonly applied filter in DMR testing is a minimum difference in methylation values. It is not uncommon to see highly significant DMR with a tiny difference in methylation values, because in general, statistical significance is dependent on both effect size and sequencing coverage. This is particularly the case for genomic intervals with high CpG coverage such as CGI. Generally, the critical value is defined as an absolute methylation difference above 0.1 or 0.2, which I found to be not necessarily appropriate in the context of tumour versus normal DMR testing.

In my exploratory analyses, I noticed that normal tissues have roughly the same genome-wide methylation values around 0.7. Let ΔM be the methylation

difference in a DMR, formally defined as $\Delta M = M_t - M_n$. M_t is bounded between $[0, 1]$. When $M_n = 0.7$:

$$0 \leq M_t \leq 1$$
$$0 \leq \Delta M + M_n \leq 1$$
$$-0.7 \leq \Delta M \leq 0.3$$

As long as $M_n \neq 0.5$, the range of ΔM is not centred around 0. Choosing a hard cutoff may either under-filter the differentially hypomethylated regions (hypoDMR), or over-filter the differentially hypermethylated regions (hyperDMR). To overcome this problem, effect size cutoffs for hypo- and hyperDMR are defined separately unless otherwise specified.

Model dispersion

Last but not least, DMRs with estimated dispersion ≥ 0.25 are also excluded, because this indicates that the samples are highly variable at these genomic loci, which may suggest the violation of model assumptions where samples within the same comparison group belong to the same distribution. This step acts as an additional filter akin to the Cook's distance diagnostic to remove likely invalid genomic loci, but is applied after the p adjustment step to avoid interfering with fdr.

3.1.4 Major limitations

While the mean observed methylation can be modelled as a linear function of tumour fraction (see Eq. (3.3)), the variance structure is much more complicated. In the *DSS* framework, the variance structure for each locus is modelled using beta-binomial distributions with the same dispersion parameter. In other words, different comparison groups are expected to have roughly the same variability given the sequencing depth is similar. This assumption may be true when comparing normal to normal, but it has been reported that tumours can have hyper-variable methylation [278]. To address this phenomenon appropriately, each sample should

ideally be modelled as a mixture of beta-binomial models with different dispersion parameters. However, doing such would drastically increase the mathematical complexity (if solvable) and computational time, which defeats the purpose of using an efficient algorithm like *DSS*. The direct consequence of using a "blanket" dispersion parameter is that, if tumour really has higher variability than non-tumour, the variance of the non-tumour methylation would be overestimated. Hence, the differential testing for some genomic interval may be underpowered.

3.2 Detecting broad methylation patterns in low tumour fraction samples

3.2.1 Rationale

DNA methylation is traditionally analysed as a quantitative assay, where methylation value is defined as the signal of methylated CpGs over all CpGs for a given genomic position. However, more recently, read-based approaches have been developed [265–269, 279, 280], which confer additional insight by utilizing short-range phased methylation. As illustrated in Fig. 3.3, read-based methods can distinguish methylation patterns even with exactly the same overall methylation. Loyfer et al. [269] demonstrated that their approach can be highly sensitive at detecting as low as 0.1% of cell type specific methylation.

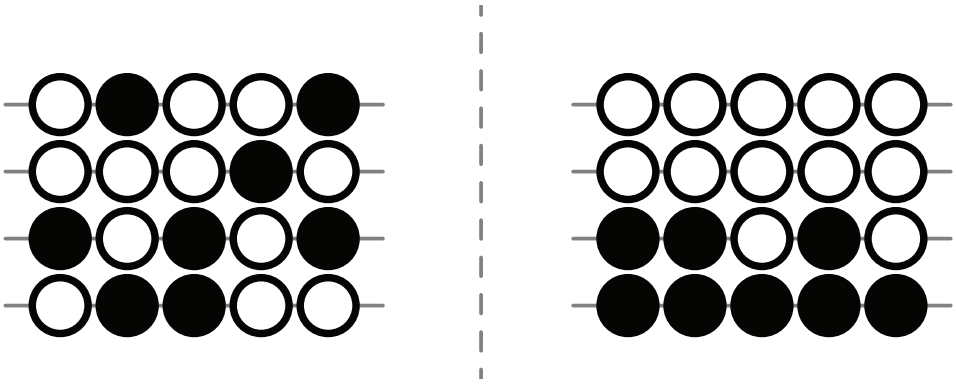


Fig. 3.3: Illustration of advantage of read based methylation. Methylation information is phased for each read, and can distinguish between the two illustrated scenarios, where overall methylation are both 40%.

Yet, most if not all existing efforts are focused on analysing methylation pattern in CpG-dense regions, partly because they are more likely to be functional, but also because only these regions have enough phased CpGs using short-read sequencing platforms. This limits the analysable region to less than 1% of the genome.

Given that global hypomethylation occurs in around 50% of the cancer genome [281], there may be a direct clinical application to have an approach to detect broad hypomethylation beyond CpG-dense regions on a read-based level, which may facilitate detecting tumour methylation patterns in cfDNA.

An additional advantage of using broad regions is that this may lower the sequencing depth requirement for accurate detection. Tumour fraction in cell-free DNA is typically below 10% and often in the 0.1% range, meaning there may only be 1 tumour DNA molecule for every 1000 cfDNA fragments. Suppose there are 1000 CpG sites that are specifically methylated in cancer, and it is a clinically feasible cost to perform 1x methylation sequencing, which means on average each tumour-specific loci would be covered by 1 read. Then, at 0.1% tumour content, only 1 tumour specific loci would capture tumour methylation, whereas the rest would capture normal cfDNA methylation. This small signal must then overcome the background noise, which may come from sequencing error, PCR bias, platform-related technical artefacts, 3' end repair defect, etc. On the other hand, if broad bins were used, each locus would be covered by a much higher number of reads, hence increasing the chance to capture at least 1 tumour DNA fragment for each locus.

3.2.2 Binomial mixture modelling

Mixture models can be used to model multimodal data distributions. Intuitively, it considers an overall distribution to be a mixture of subpopulations, and makes inference on their properties such as the means and mixing proportions. One could either pre-specify a range of mixing components in the case of finite mixture modelling, then choose the best number of components according to model

performance metrics such as AIC or BIC, or use a parameter-free approach in the case of infinite mixture modelling.

Each sequencing read must come from either tumour or non-tumour cells and not both. Suppose for a 1 Mb region, tumour has methylation value of M_t whereas non-tumour has M_n. Let t denote the cell of origin, with non-tumour as 0 and tumour as 1. We may express the read-based methylation M_r as follows:

$$M_r|t = 0 \sim Bin(n, M_n), \text{and}$$
$$M_r|t = 1 \sim Bin(n, M_t)$$

Where Bin denotes binomial distribution, and n is the number of CpG present on the read. The distribution of the overall mixture M is a weighted sum of its components, and is expressed as:

$$M = P(t = 0)Bin(n, M_n) + P(t = 1)Bin(n, M_t)$$

Where $P(t = 0)$ and $P(t = 1)$ represents the probability of cell of origin. The model coefficients can then be estimated subsequently using expectation-maximization (EM) algorithms.

Mixture modelling has been implemented in many statistical tools, and finite mixture modelling with a range of components from 1 to 3 has been chosen over infinite mixtures due to runtime considerations. The tool used in this work is the R package `flexmix`.

3.2.3 Model performance using simulated data

To assess the potential of this method, the model performance is first tested against simulated data with known ground truth parameters. Detailed codes can be found in Appendix A.2. In summary, 15 simulations were run for each pre-defined tumour content levels, and the mixture model can detect between 1% to 10% tumour content with $1e4$ sequencing reads at a methylation difference of 0.6, and down to 0.1% tumour content with $1e5$ sequencing reads, as shown in

Fig. 3.4. No false positives have been detected in the limited number of simulations. Having a very low false positive rate is critical to any screening tests of low-prevalence diseases, such as cancer.

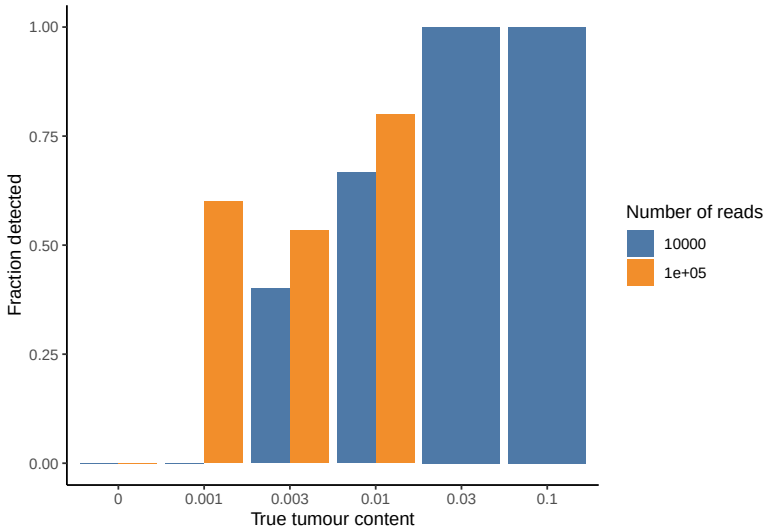


Fig. 3.4: Chance of detecting tumour reads. 15 simulations were run for each tumour content levels, at $1e4$ (blue) and $1e5$ (orange) sequencing reads per hypothetical loci. A locus is defined as "positive" if more than 1 mixture component is detected. No tumour was detected at 0.1% with $1e4$ reads, but 9 out of 15 were detected with $1e5$ reads. No false positive was detected in this simulation.

However, it is also noted that for low tumour content, mixture modelling consistently underestimates the tumour content, shown in Fig. 3.5.

The performance on real data may be a lot worse due to the presence of noise and technical errors. Therefore, careful selection of tumour-specific loci to minimize background noise is warranted.

3.2.4 Selection of tumour-specific loci

In the context of tumour detection from cfDNA, the major task is whether we can detect reads coming from tumour versus those coming from normal cells, which are mainly immune cells, in the selected genomic loci. Conceptually, we can maximize the tumour signal by choosing genomic regions that are ubiquitously (un)methylated in normal blood cells.

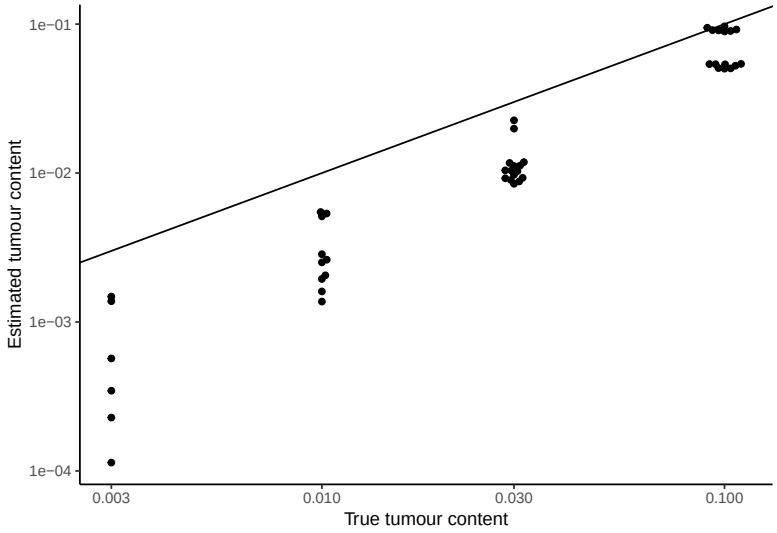


Fig. 3.5: Beeswarm plot of tumour content by binomial mixture modelling at $1e4$ reads against ground truth. Straight line represents $y = x$. The estimated tumour content is consistently below the line of equivalence, especially for tumour content below 10%.

To obtain the tumour-specific loci, a tumour versus non-tumour DMR is first performed using the framework described in Section 3.1 on 2 kb windows constructed around CpG islands. Then, circular binary segmentation (CBS) from `DNAcopy` [282], an R package commonly used in copy number calling, is repurposed and performed on the methylation difference to obtain broad intervals of methylation change. Coverage for each bin is divided by 1000 and used as weights in the CBS algorithm, which greatly reduced over-segmentation under default settings. A filtering step is then performed on the segmented broad DMRs, which requires a segment to have at least 40% of genomic intervals to have $fdr < 0.05$, dispersion < 0.1, and a mean methylation difference < -0.4 or > 0.2. The rationale behind the asymmetric size effect cutoffs has been explained in Section 3.1.3, and the values of the cutoffs are defined according to visual exploration of the data.

Subsequently, a second filtering step is performed by using TAPS data of normal cell types, including B cells, T cells, Natural Killer (NK) cells, monocytes, neutrophils, and eosinophils. Read-based methylation is calculated for each cell type and then mixed together to perform binomial mixture modelling on the filtered broad DMRs from the previous step. Initially, loci with only 1 mixture

component were chosen for further analyses. However, it appeared later on that the number of loci using this strict filter is too few for low sequencing depth data, and therefore a wider set of loci was defined, which includes also loci that have more than 1 mixture component as long as the largest methylation difference among the components is ≤ 0.65.

Normal cell TAPS data were taken with permission from unpublished work of M Inoue from C Song's lab. No healthy control nor patient cfDNA were used in defining tumour-specific loci. Results of the LUD2015-005 trial cfDNA data analysis will be presented in Chapter 4.

3.3 Tumour fraction and copy number aware subtyping

3.3.1 Motivation

It has been demonstrated above that the DMR framework can handle more than 1 covariate, such as a grouping variable like tumour subtype, or clinical outcome. However, it remains a question of how the subgroups should be defined. In our case, we are particularly interested in defining methylation subtypes of OAC. There have been several attempts to this goal [8, 231, 234, 235, 283], but none of them formally addressed the issue of tumour fraction.

To demonstrate the importance of considering tumour fraction in clustering, I re-analyzed the data presented by Liu et al. [235], which inferred the existence of several CpG island methylator phenotypes (CIMPs) in gastrointestinal adenocarcinomas, including CIMP from Epstein-Barr virus (EBV) positive gastric cancers (EBV-CIMP), CIMP high (CIMP-H), CIMP low (CIMP-L), and non-CIMP. CIMP-L samples are further classified into gastroesophageal (GEA-CIMP-L) and colorectal (CRC-CIMP-L) subtypes. However, instead of downloading the processed data from Liu et al., the version from The Cancer Genome Atlas (TCGA) data portal was used.

In the original analysis, tumour clusters were defined by performing hierarchical clustering on dichotomized methylation values using a cutoff value of 0.3 in loci

that are unmethylated in normal tissues. The rationale is that any methylation in these selected loci is likely contributed by the tumour, and dichotomizing the methylation value can help to handle variations brought about by tumour fraction. However, this approach of clustering is in fact still highly dependent on the arbitrary cutoff and the distribution of sample tumour fraction.

Figure C.1a shows recreation of the methylation heatmap and clustering results using the original 0.3 cutoff. The clusters are then annotated by the published cluster annotation in Liu et al. with the largest sample overlap. The clusters roughly agree with the annotation presented in [235], although complete correspondence could not be achieved, probably due differences in data processing pipeline. A comparison between the published cluster annotation and the recreated cluster annotation can be found at Fig. C.2.

Figure C.1b shows a similar plot using a new cutoff of 0.28. The change in cluster membership is shown in Fig. 3.6, showing that adjusting the cutoff by merely 0.02 already changed the cluster memberships of 21.4% samples.

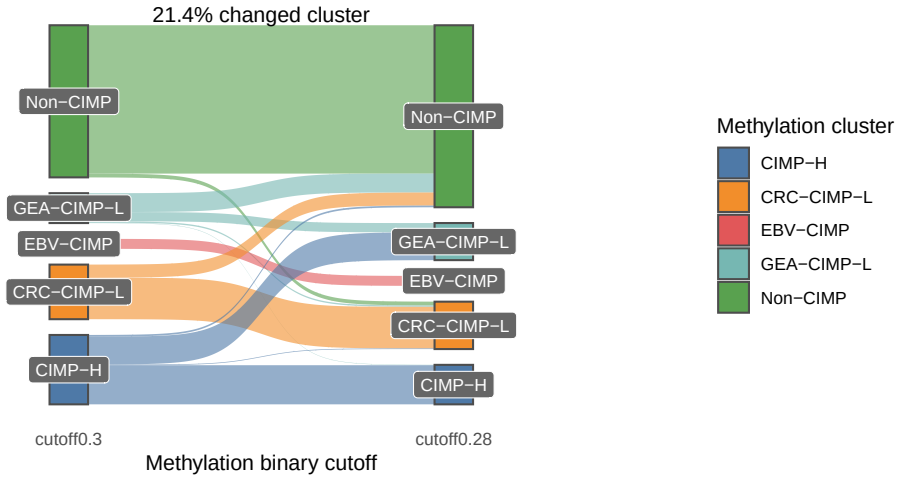


Fig. 3.6: Sankey plot to visualize the effect of cutoff used to dichotomize methylation values on methylation clusters. A minute change of 0.02 leads to changes of cluster in 21.4% of samples.

Similarly, dilution of samples also causes a shift in cluster membership, shown in Fig. 3.7. Methylation data of CIMP-H samples were diluted with the mean methylation of non-cancer control samples per loci to 80%, 60%, and 40% respectively.

The diluted samples are subsequently re-clustered with the original tumour data using 0.3 cutoff. Dilution causes CIMP-H samples to cluster with GEA-CIMP-L and non-CIMP samples, as more and more loci fall beneath the dichotomizing cutoff. As the dilution approaches the dichotomizing cutoff, a dramatic shift towards CIMP-L and non-CIMP phenotypes is observed. Interestingly, diluted samples are preferentially classified into the GEA-CIMP-L subtype instead of CRC-CIMP-L subtype, despite around 30% of CIMP-H being CRC samples. A breakdown of cluster shift upon dilution by site of cancer can be found at Fig. C.3.

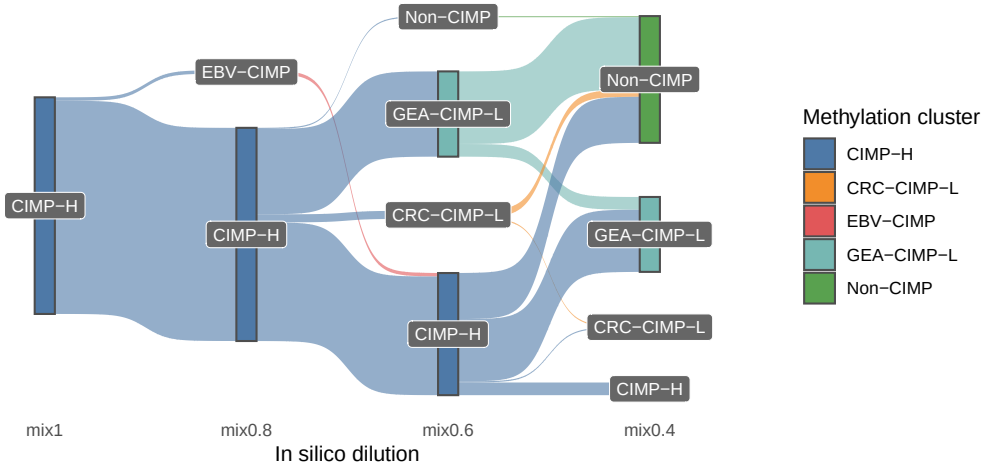


Fig. 3.7: Sankey plot to visualize the effect of tumour fraction on methylation cluster. Methylation data of CIMP-H samples were diluted with the mean methylation of non-cancer control samples per loci to 80%, 60%, and 40% respectively. The diluted samples are subsequently re-clustered with the original tumour data, and re-annotated as described before.

Taken together, the clustering results in Liu et al. might have suffered from uneven tumour purity, where some GEA-CIMP-L and non-CIMP may be contributed by CIMP-H samples with lower tumour contents.

Zhang et al. raised the importance of accounting for tumour purity when clustering methylation array data and developed *InfiniumClust* [284]. The algorithm in the publicly available deposit only considers tumour purity, but can be easily modified to consider copy number as well by replacing tumour purity with a vector of local tumour fraction according to Eq. (3.2). However, despite the potential ability to be copy number aware, there are 3 major reasons why this algorithm may not be

suitable for our data. Firstly, *InfiniumClust* only cluster the top 100 methylation loci with the highest variance. Incrementing this number greatly increases the computational time. This may be desirable for array type data, but may not be enough to make full use of genome-wide data. Secondly, TAPS data has a discrete distribution, whereas *InfiniumClust* takes methylation as a continuous variable bounded between 0 and 1. Hence, one of the key assumptions is violated. Lastly, *InfiniumClust* models the non-tumour methylation as a single distribution, which assumes that all tumour subtypes have the same microenvironment composition. However, multiple studies have suggested that there may be clinically relevant differences in the immune microenvironment of OAC [8, 237, 283], and therefore a clustering method that allows differences in non-tumour methylation would be more appropriate.

3.3.2 Tumour fraction and copy number aware dichotomization of methylation

DNA methylation is a binary event, and it makes sense to report certain loci as being methylated or not, rather than a variable between 0 and 1. Following the spirit of [231, 235], it is reasonable to classify subtypes based on matrices of digital variables.

However, instead of dichotomizing methylation by a single cutoff, I developed 2 model-based approaches that can be used to binarize methylation in consideration of tumour DNA fraction. Detailed code is available at Appendix A.3.

Binomial test. The first approach starts with building a beta-binomial regression model with tumour DNA fraction as covariate, which is identical to that in Fig. 3.1. Next, based on the regression model coefficients, the predicted methylation for each tumour at the corresponding tumour purity is calculated, assuming there is no subtype. Then, a null hypothesis can be assumed, that if there is no subtype, then sample methylation would follow a binomial distribution with the expected mean being the predicted mean from the model. Binomial test can then be performed to see how likely it is to obtain the methylation sequencing counts given the predicted

methylation. Using a p value cutoff of 0.05, each data point can then be classified as significantly above, below, or not significantly different from the overall mean.

Figure 3.8 shows the performance of this approach in simulated data, with ground truths as follows:

Subtype	Non-tumour methylation	Tumour methylation
A	0.8	0.1
B	0.6	0.8

Table 3.1: Ground truth methylation values.

Out of 30 samples, 14 samples were classified to be significantly different from the overall mean, and 12 samples were classified correctly in the simulated example.

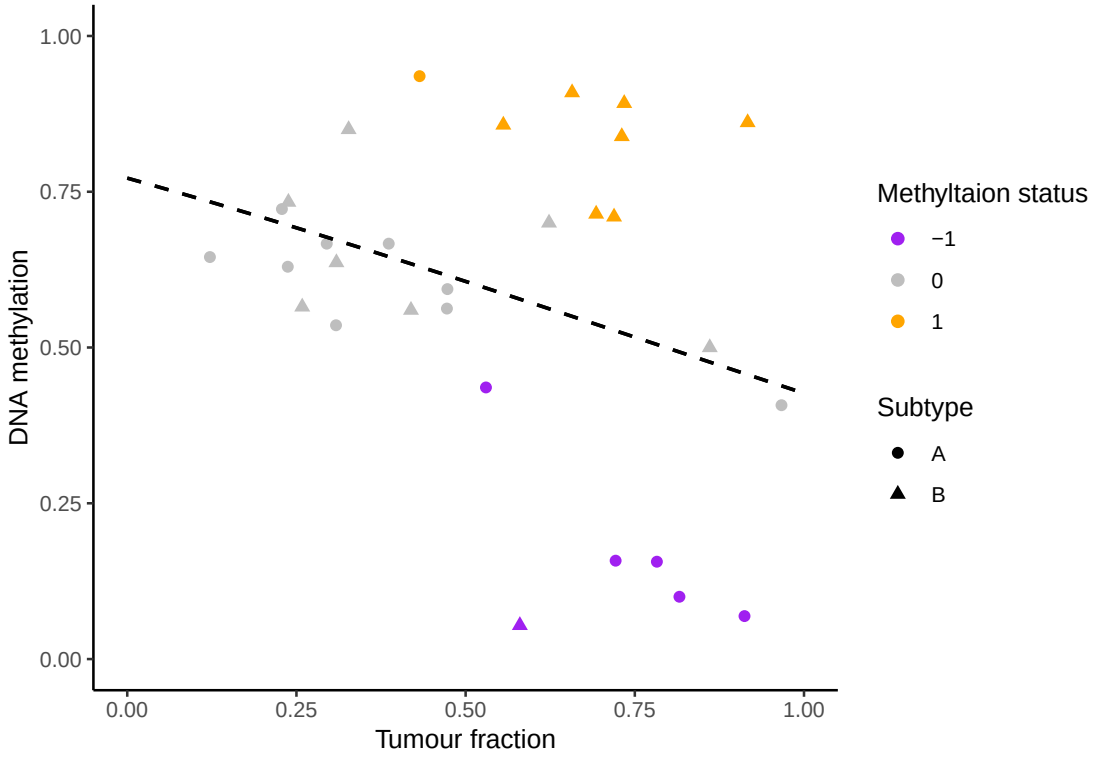


Fig. 3.8: Shape of the dot corresponds to the ground truth tumour subtype, whereas colour of the dot represents the subtype prediction. Grey dots represent samples that are not statistically different from the predicted methylation. The black dotted line represents the overall mean, as modelled with only tumour fraction as covariate.

However, the binomial test approach may lead to inaccurate results in scenarios where tumour methylation of subtype A is higher than subtype B, and non-tumour methylation of subtype B is higher than subtype A. High purity subtype A samples would be classified as higher than overall mean, whereas low purity subtype A samples would be classified as lower than overall mean. Thus, a second approach is developed.

Mixture modelling. The alternative is to build a finite mixture model using binomial regression of n components. This approach is similar in principle to that used in Section 3.2, except tumour DNA fraction is included as a covariate. Since we are mostly interested in the binary methylation status, we choose component $n \leq 2$. By default, mixture modelling would assign a cluster to each sample, even in ambiguous cases. This can be avoided by implementing a filter to the posterior probability of the cluster assignment. Only cluster assignment with a posterior probability of ≥ 0.95 will be accepted in the final output.

Figure 3.9 shows the performance of this approach in the same simulated data as above. Out of 30 samples, the mixture modelling approach classified 18 samples, and 15 samples were classified correctly.

The above procedures are repeated for every loci genome wide. Then, principal component analysis (PCA) is performed to reduce the dimension, and hierarchical clustering is carried out using the Ward D's method.

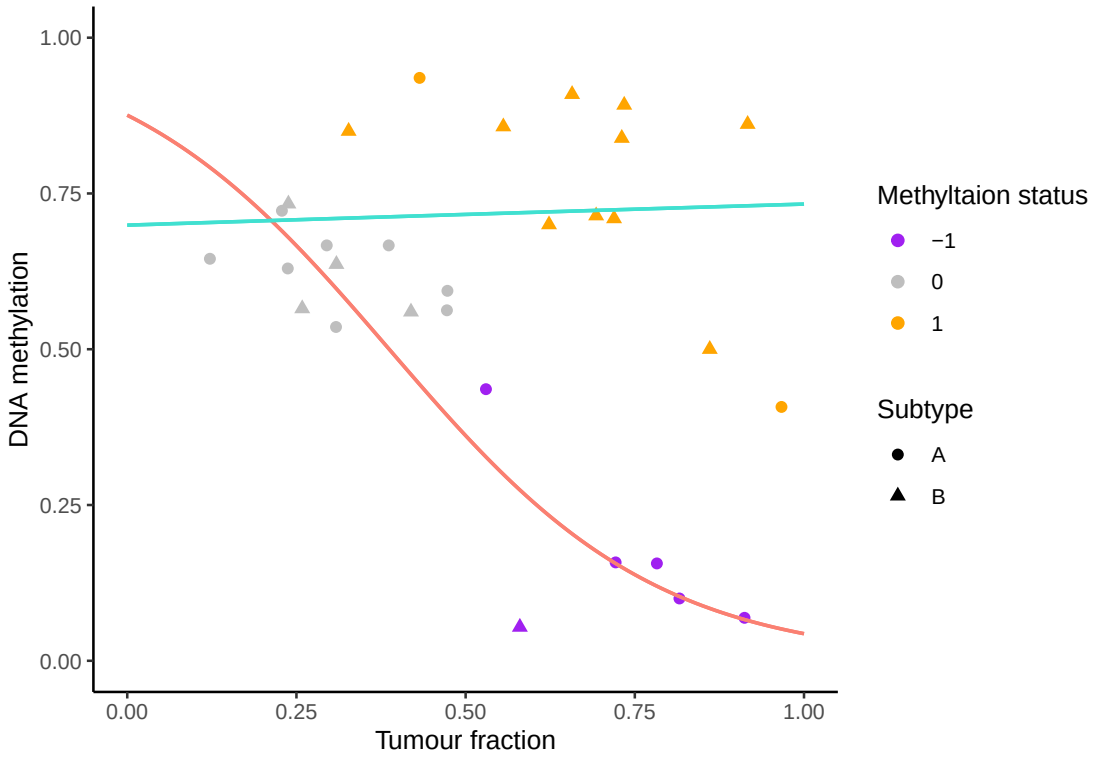


Fig. 3.9: Shape of the dot corresponds to the ground truth tumour subtype, whereas colour of the dot represents the subtype prediction. Grey dots represent samples with a scaled posterior probability < 0.95. The salmon and turquoise lines represents the two mixing components modelled by the algorithm.

3.3.3 Assessing model performance

Apart from obtaining a stable, robust data-driven clustering, it is also important to assess the evidence of whether the subtyping is valid. The gold standard would of course be to experimentally validate the hypothesis, or cross-validate with an independent cohort. However, it is also possible to computationally assess whether introducing additional covariates in the model design can explain the dependent variables better. Using the example illustrated in Fig. 3.2, we may formulate our question as such:

$$\text{Does model 1}: \quad M_o = \beta_0 + \beta_1 \lambda_i$$

$$\text{or model 2}: \quad M_o = \beta_0 + \beta_1 \lambda + \beta_2 S + \beta_3 \lambda S$$

fit the data better?

This is known as the model-selection problem. Commonly used methods are the Akaike's Information Criteria (AIC) and Bayesian Information Criteria (BIC), which applies certain penalty to the log-likelihood of the fitted model to get a score. The model with the least AIC or BIC would be considered best.

I implemented the calculations of AIC and BIC in line with those defined in the `stats` package in R. Their definitions are:

$$AIC = -2ln(L) + 2k$$
$$BIC = -2ln(L) + ln(D)k$$

Where $ln(L)$ is the log-likelihood, k is the degree of freedom or number of fitted model parameters **including the variance**, and D is the number of data points. Both AIC and BIC penalize on the number of parameters being fitted, but BIC also varies its penalty term based on sample size instead of a fixed number. Despite the differences, both criteria have similar rationale of choosing the simplest model that can best explain the data, which requires striking a balance between choosing the best predictors and overfitting the parameters.

For weighted least squares of D samples, the log-likelihood is defined in R as:

$$ln(L) = \frac{1}{2}\sum_{d=1}^{D} ln(\omega_d) - \frac{D}{2}\left[ln(2\pi) + 1 - ln(D) + ln(\boldsymbol{e}^T\boldsymbol{e})\right]$$

Where ω_d is the weight of the d$^{\text{th}}$sample, which is defined as the diagonal of inverse variance matrix $\boldsymbol{V}^{-1}$, and $\boldsymbol{e}^T\boldsymbol{e}$ is the squared sum of weighted residuals. Detailed definition of $\boldsymbol{V}$ and $\boldsymbol{e}$ can be found in Eqs. (B.1) and (B.4) of Appendix B.

Consider a dataset of D samples and N genomic loci. The total number of data points equals $\sum_{i=1}^{N} D_i$ where D_i is the number of non-missing data points for each locus. For each locus, k parameters are estimated. Therefore, the total number of fitted parameters equals $N \times k$, where $k = 3$ for model 1 and $k = 5$ for model 2. The joint log-likelihood is just the sum of log-likelihood across all

N loci. Thus, the formula for the total AIC is:

$$\begin{aligned} AIC_{total} &= -2\sum_{i=1}^{N} ln(L_i) + 2Nk \\ &= \sum_{i=1}^{N} [-2ln(L_i) + 2k] \\ &= \sum_{i=1}^{N} AIC_i \end{aligned}$$

Where AIC_i is the AIC of the i$^{\text{th}}$loci. Similarly, the formula for BIC is:

$$\begin{aligned} BIC_{total} &= -2\sum_{i=1}^{N} ln(L_i) + ln\left(\sum_{i=1}^{N} D_i\right) Nk \\ &= -2\sum_{i=1}^{N} ln(L_i) + \sum_{i=1}^{N} ln(D_i)k - \sum_{i=1}^{N} ln(D_i)k + Nln\left(\sum_{i=1}^{N} D_i\right) k \\ &= \sum_{i=1}^{N} [-2ln(L_i) + ln(D)k] + \left[Nln\left(\sum_{i=1}^{N} D_i\right) - \sum_{i=1}^{N} ln(D_i)\right] k \\ &= \sum_{i=1}^{N} BIC_i + \left[Nln\left(\sum_{i=1}^{N} D_i\right) - \sum_{i=1}^{N} ln(D_i)\right] k \end{aligned}$$

Assume missing data are rare, i.e. $D_i \approx D$. Then:

$$\begin{aligned} BIC_{total} &\approx \sum_{i=1}^{N} BIC_i + [Nln(N \times D) - Nln(D)]\, k \\ &\approx \sum_{i=1}^{N} BIC_i + Nln(N)k \end{aligned}$$

It can be seen that while the total AIC is the sum of AIC of each locus, the total BIC has an additional penalty term that gets very heavy in a genome-wide dataset, raising the concern of whether its use is appropriate. Unfortunately, this question is beyond the scope of this thesis. Nonetheless, BIC is still calculated for each fitted model design for reference, but AIC is used primarily to determine which is the better model.

4

Tumour associated methylation changes

Contents

4.1 Overview

When testing for tumour-specific DMR, a common question is what tissue type should be used as the non-tumour control. "Adjacent normal", defined as histologically normal tissues adjacent to the tumour, is generally an acceptable option. Yet, OAC occurs at the oesophagogastric junction, and it is debatable whether oesophagus or gastric cardia should be used as the normal reference. The tumour fraction aware differential testing framework described in Section 3.1 is capable of calling tumour-specific DMR in the absence of non-tumour samples, thus circumventing the problem. However, it does create new problems such as

assuming the cell composition of the non-tumour component are the same across all samples, which is likely false. Fortunately, lineage-specific methylation marks are rare on a genome-wide scale, and the majority of the normal genome has ubiquitous methylation levels. Therefore, although the underlying cell composition may be different, the non-tumour component of each sample has largely the same methylation. The methylation in lineage-specific loci may be different across samples, but these regions would have high dispersion values and may be filtered out by the Cook's distance diagnostic.

In this chapter, I will explore the tumour-associated methylation changes compared to the non-tumour compartment within the tumour microenvironment. I will systematically describe the changes in tumour DNA methylation with respect to different genomic elements such as histone marks, regulatory elements, and relate methylation changes to RNA transcript abundance. Lastly, I will also demonstrate the possible clinical applications of tumour-specific methylation.

4.2 Results

4.2.1 Systematic description of tumour-associated methylation changes

Global methylation landscape of tumour and non-tumour compartments

The average tumour and non-tumour methylation at each genomic locus are obtained from the beta-regression model by substituting tumour fraction λ in Eq. (3.3) with 1 and 0 respectively, as illustrated previously in Fig. 3.1. Their global methylation landscapes are then aggregated and compared in Fig. 4.1.

Compared to non-tumour, OAC demonstrates drastically different methylation profile, consistent with the known pattern of global hypomethylation and focal hypermethylation. Although the aberrant methylation pattern is well recognized, the magnitude of such change is seldom described. By correcting for tumour fraction, it can be seen that there is a full range of hypomethylation going down from 0.8 to almost 0, whereas the range of hypermethylation goes up from 0 to

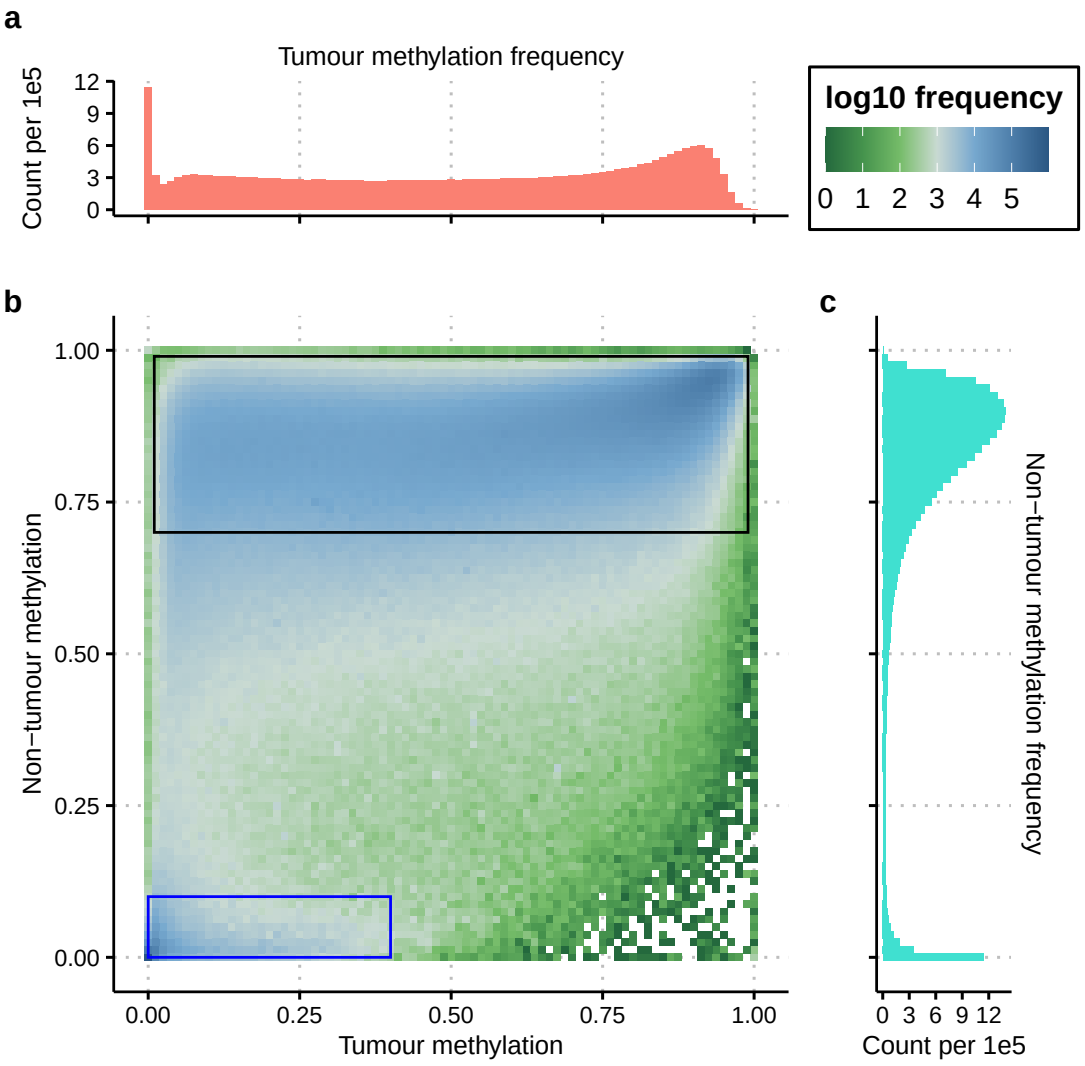


Fig. 4.1: Global methylation landscape of tumour and non-tumour. (**a,c**) Histograms of tumour (salmon) and non-tumour (turquoise) methylation, with scales aligned to the 2D density plot. Histogram is constructed using per-CpG methylation averaged over each genomic bin defined in Section 3.1.2. (**b**) 2D density plot of tumour versus non-tumour methylation. The frequency is coloured on a log-scale, with a green-white-blue transition from low to high. Tumour has both hypomethylation of highly methylated regions (0.7 to 1, black box) and hypermethylation of lowly methylated regions (0 to 0.1, blue box).

only about 0.4 according to the TAPS assay. The plot roughly agrees with that reported in colorectal cancer [207].

For comparisons of methylation landscape between the estimated non-tumour and the measured adjacent normal of the trial patients, which were not used in building the DMR model, see Figs. C.4 and C.5.

Late-stage OAC has near-complete hypomethylation in late replicating domains

It has been recently characterized that global hypomethylation in tumour is non-random, organized in blocks of million base-pair (Mb) scale, and enriched in heterochromatin regions [205, 207]. The proposed mechanism is the passive loss of CpG methylation due to incomplete inheritance during DNA replication, driven by excessive cell division [209, 285]. DNA replication is highly regulated and happens in a defined temporal order [286]. Chromatin in late replicating domain (RepD), which largely coincide with heterochromatin, has the least time to copy methylation marks from the template to the nascent strand, and is therefore especially prone to demethylation. This phenomenon is present in non-tumour cell types as well with a modest decrease in methylation, hence named partially methylated domains [287].

Replication domain annotation is downloaded from [288], and methylation is plotted for constitutive early and late RepDs in Fig. 4.2a. 50% of CpGs in late RepDs have less than 0.31 methylation, with the mode at around 0.1. Fig. 4.2b shows the methylation in a spatial context, demonstrating the drastic decrease of tumour methylation in late RepDs in the scale of 10 Mb.

Repeat element is not enriched for hypomethylation in OAC

There is good evidence that repeat element (RE) hypomethylation and reactivation is present in a pan-cancer fashion, summarized in [289], and may play a role in carcinogenesis by inducing genome instability [290, 291]. In particular, LINE-1 hypomethylation has been shown to be prognostic in gastric cancer and proposed to be a biomarker [292]. While there is an established link between RE hypomethylation and cancer development, whether there is a mechanism to specifically demethylate RE remains unknown.

To investigate this, the methylation data is annotated for RE using the UCSC RepeatMasker Track. Genomic intervals with non-tumour methylation > 0.65 are selected, and loci with $fdr < 0.01$ and lower methylation in tumour are defined as significant tumour hypoDMR. The proportion of RE hypoDMR is compared to that

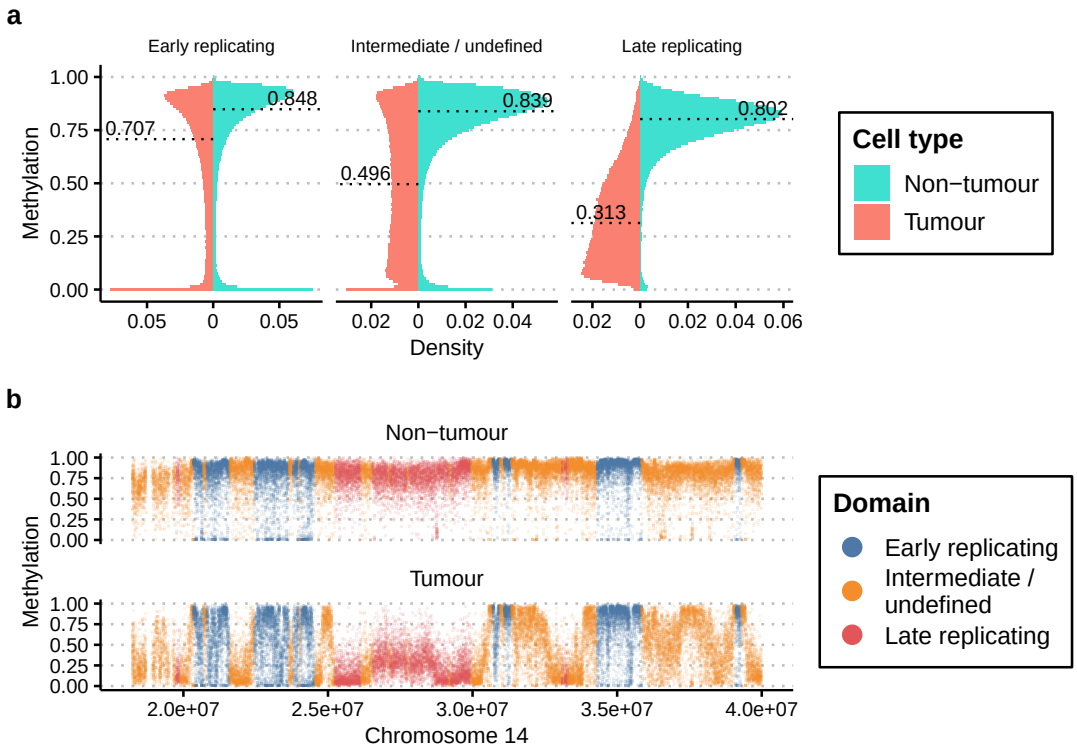
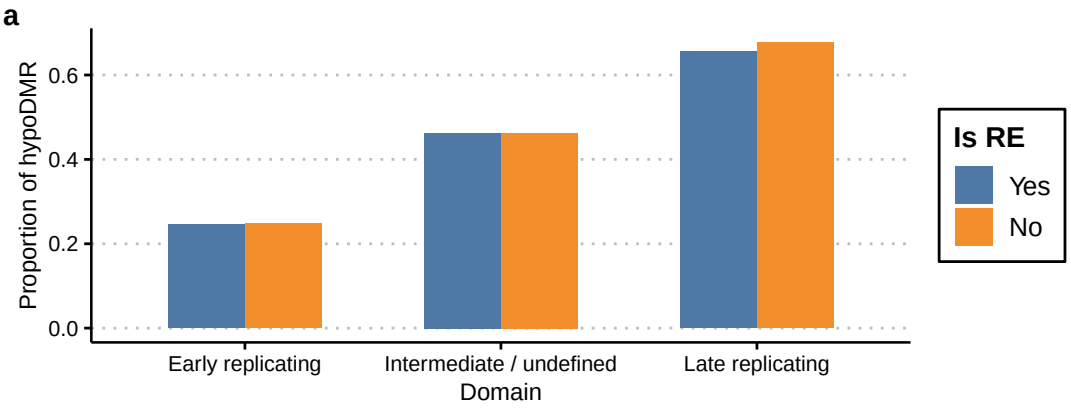


Fig. 4.2: Replication domain methylation of tumour and non-tumour. (**a**) Mirrored histograms of replication domain methylation of tumour and non-tumour, with density on x-axis and methylation on y-axis. Median methylation for each group is marked with a dotted line and annotated. Tumour is labelled with salmon colour and non-tumour with turquoise. There is a trend of decreasing median methylation from early to late RepDs in both tumour and non-tumour, with a bigger magnitude in tumour. (**b**) Visualization of tumour and non-tumour methylation in chromosome 14. Each dot represents a genomic interval with RepD annotated in colors.

of non-RE hypoDMR using binomial regression with replication domain included as covariate, shown in Fig. 4.3. In early and intermediate RepDs, RE has similar chance to be hypomethylated in tumour compared to non-RE. In late RepDs, RE is less likely to be hypomethylated with an odds ratio (OR) of 0.91. In other words, the hypomethylation of RE is well-explained by the effect of passive demethylation due to replication stress. It is unlikely that there is a pathway to specifically demethylated RE. On the contrary, the slight depletion of RE hypoDMR in late RepD suggests that there might actually be a mechanism to prevent demethylation of RE. Alternatively, cancer clones with too much RE demethylation may be selected out due to increased immunogenicity [293]. For the global methylation landscape

of different classes of REs in different RepDs, see Fig. C.7.



b

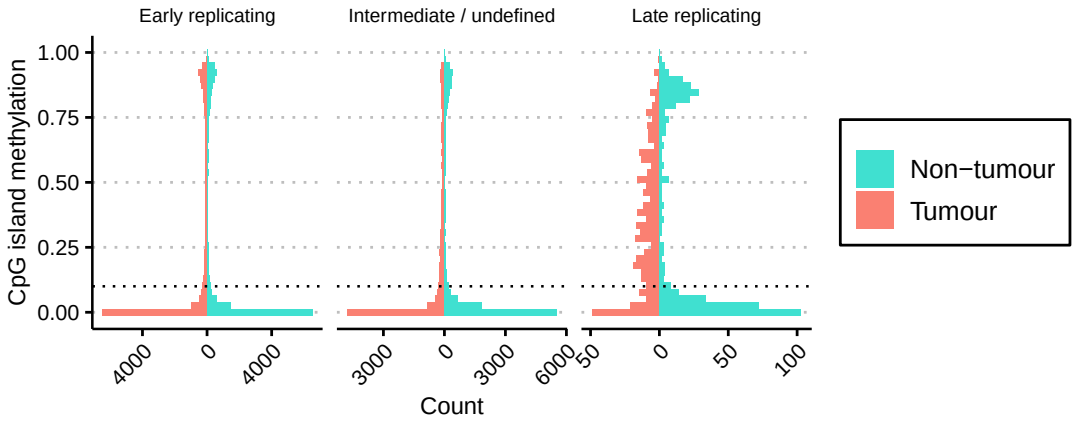
Domain	Is RE	hypoDMR	Valid intervals	Odds	Odds ratio	p value
Early replicating	No	114222	460367	0.330	(ref)	
	Yes	150953	613037	0.327	0.99	2.60e−02
Intermediate / undefined	No	607841	1314373	0.860	(ref)	
	Yes	663045	1433666	0.860	1.00	9.67e−01
Late replicating	No	141521	209090	2.090	(ref)	
	Yes	171365	261315	1.910	0.91	3.10e−52

Fig. 4.3: RE not enriched for hypomethylation in OAC. (**a**) Barchart showing the proportion of the number of hypoDMR over the total number of genomic intervals with non-tumour methylation > 0.65 in RE and non-RE respectively. (**b**) Table describing the number of hypoDMR with odds ratio calculated for each RepD using non-RE as reference. Binomial regression is performed with the count of hypoDMR and non-hypoDMR as dependent variables, and the interaction between RE status and replication domain as independent variables. p values are obtained by model contrast of RE status in each replication domain. Acronym: ref, *reference*.

HyperDMR are enriched in late replicating domains

An additional observation in Fig. 4.2 is the depletion of fully unmethylated CpGs in late RepDs for both tumour and non-tumour. Unlike the classical bimodally distributed methylation values in early RepDs, the methylation in late RepDs is largely unimodal, which is not an artefact of model estimation, and is observed in adjacent normal (Fig. C.6). Despite the relative depletion, there were suggestions from previous literature that focal hypermethylation is actually enriched in hypomethylated domains [207, 278]. To test this, CGIs with non-tumour methylation

< 0.1 are selected, and loci with $fdr < 0.01$ and higher methylation in tumour are defined as significant tumour hyperDMR. A contingency table is constructed, comparing the proportion of CGI hyperDMR in different RepDs. 60% of lowly methylated CGI became hypermethylated in late RepD, versus 13% in early RepD. Cochran-Armitage test for trend is performed and indeed shows strong enrichment of CGI hypermethylation in late RepDs (Fig. 4.4), despite the absolute number of hyperDMR being much lower.



Domain	HyperDMR in tumour	Total valid CGI	Proportion	p value
Early replicating	1196	8915	0.134	
Intermediate / undefined	2339	8433	0.277	
Late replicating	135	224	0.603	1.15e−156

Fig. 4.4: Mirrored histograms of CGI methylation of tumour (salmon colour) and non-tumour (turquoise) in different replicating domains, with methylation on y-axis and counts on x-axis. CGIs with ≤ 0.1 non-tumour methylation are identified in each domain, and is labelled as hyperDMR if the DMR test has $fdr < 0.01$ and higher methylation in tumour. A contingency table is constructed for hyperDMR CGI in each domain, and a Cochran-Armitage test is performed for the null hypothesis of there being no monotonic relationship between the proportion of hyperDMR CGI from early to late replicating domain, which is rejected with a p value of $1.15e{-}156$.

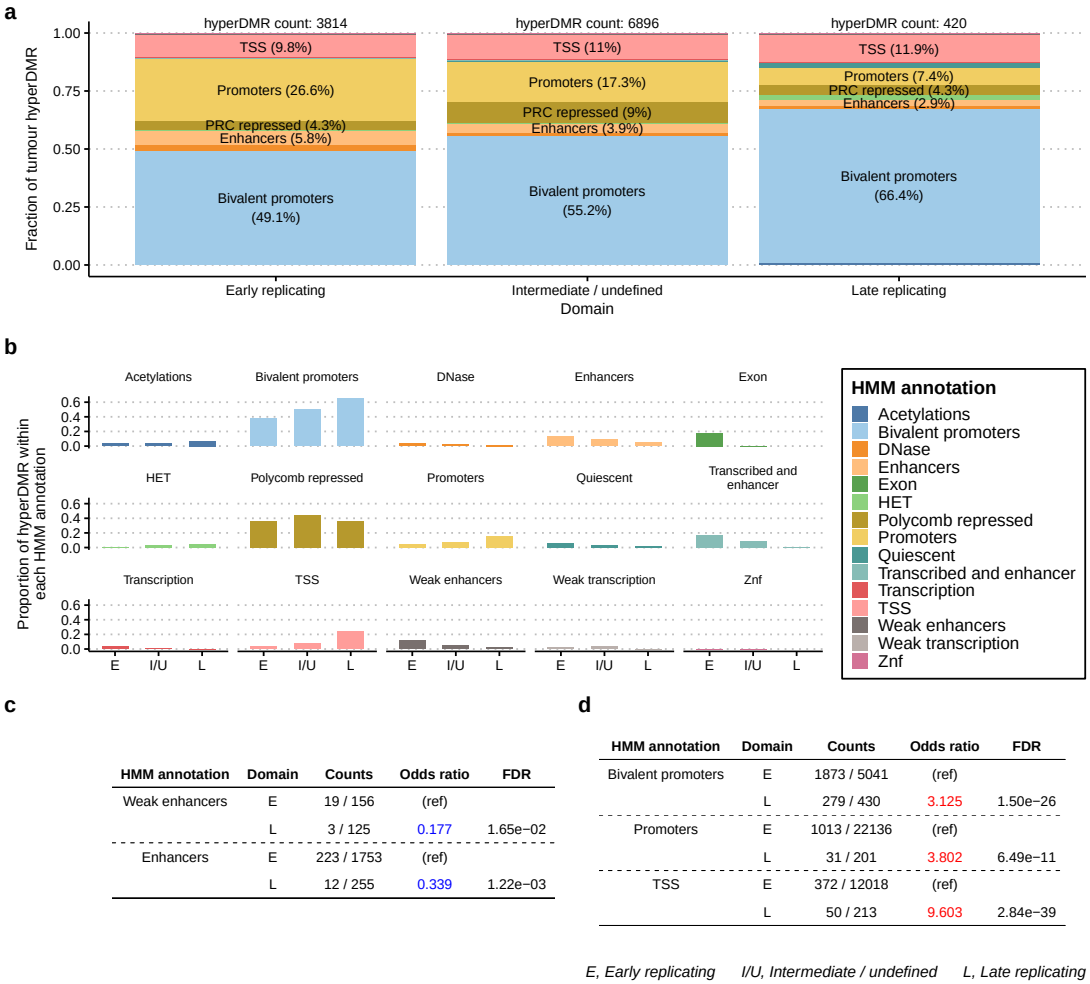
Replication domain enrichment of HyperDMR is explained by promoter methylation

The modelled tumour and non-tumour methylation so far replicates findings from existing literature quite well. However, it remains intriguing why some CpG within late RepD is hypermethylated. For further unbiased assessment of the regions with hypermethylation in tumour, the genome is annotated into 15 functional groups as defined in [275] using *ChromHMM* (henceforth referred to as HMM annotations), with gap artefacts excluded from analysis. Significant tumour hyperDMR is defined as above. Figure 4.5a shows that more than half of the hyperDMRs are bivalent promoters, and together with promoters and transcription start sites (TSSs) account for more than 80% of hyperDMR, whereas the same analysis done using CGI definition only explains about 40% (data not shown). Figure 4.5b shows the relationship between the chance of being hypermethylated and RepD for each functional group. Binomial regression is performed for test statistics and results with $fdr < 0.05$ are shown in Figs. 4.5c and 4.5d. Bivalent promoters, promoters, and TSS are more likely to be hypermethylated in late RepD, whereas enhancers, weak enhancers, and quiescent chromatin are more likely to be hypermethylated in early RepD. The remaining functional group do not have significant changes in the proportion of tumour hyperDMR in relation to RepD.

After adjusting for promoters and TSSs, late RepD is no longer associated with increased hyperDMR, suggesting that lowly methylated non-promoter and non-TSS regions do not gain significant methylation in late RepD (Fig. C.8).

Dissecting the mechanism of focal hypermethylation in cancer

Different HMM annotations are marked by various combinations of histone marks. To get a mechanistic hint of why bivalent promoters are enriched in hypermethylation, processed histone ChIP-seq peak calls are downloaded from the Roadmap Consortium [294] for digestive cell types. Peaks with $fdr < 0.01$ and overlapped by at least 2 datasets are taken as a consensus set. A similar procedure done in the above section is performed using histone marks and CGI, instead of HMM annotation, to



c

HMM annotation	Domain	Counts	Odds ratio	FDR
Weak enhancers	E	19 / 156	(ref)	
	L	3 / 125	0.177	1.65e−02
Enhancers	E	223 / 1753	(ref)	
	L	12 / 255	0.339	1.22e−03

d

HMM annotation	Domain	Counts	Odds ratio	FDR
Bivalent promoters	E	1873 / 5041	(ref)	
	L	279 / 430	3.125	1.50e−26
Promoters	E	1013 / 22136	(ref)	
	L	31 / 201	3.802	6.49e−11
TSS	E	372 / 12018	(ref)	
	L	50 / 213	9.603	2.84e−39

E, Early replicating *I/U, Intermediate / undefined* *L, Late replicating*

Fig. 4.5: Characterizing tumour hypermethylation by HMM annotations. (**a**) Barchart showing the fraction of hyperDMR that belongs to a certain functional type for each RepD. The top 5 abundant functional groups are labelled with their respective fraction. The total number of hyperDMR in the respective RepD is labelled at the top of each bar. (**b**) Barchart showing the proportion of tumour hyperDMR in valid loci within each HMM annotation for different RepDs. (**c,d**) Table describing the counts of hyperDMR and non-hyperDMR, odds ratio, and statistics in the corresponding functional group using early RepD as reference. Increased odds is highlighted in red, and decreased odds is highlighted in blue.

see if any histone marks are independently associated with an increased chance of hyperDMR in late RepD, compared to the background enrichment of hyperDMR in late versus early RepD. Different histone marks and CGI can coexist, but for model simplicity as an exploratory analysis, it is assumed that there is no interaction between them and the effect of each histone mark is independent.

Late vs early replicating with	Odds ratio	SE	FDR
No histone mod / CGI	1.8 (ref)	0.0763	7.13e−14
CGI	4.73	0.1320	7.63e−31
H3K27me3	2.48	0.4010	4.22e−02
H3K4me1	6.26	0.7090	1.92e−02
H3K9me3	0.157	0.6370	8.09e−03
H3K27ac	0.527	0.4740	2.26e−01
H3K36me3	1.31	0.4450	5.75e−01
H3K4me3	0.52	0.5650	2.96e−01
H3K9ac	1.04	0.8180	9.62e−01

Fig. 4.6: Effect of histone on the odds of being hypermethylated. The effect of each histone and CGI is assumed to be independent, but varies according to the RepD, as established above. The first term shows the background enrichment of hyperDMR comparing late versus early RepD. The remaining terms are the contrast between the background and the enrichment of hyperDMR in late RepD when the specified histone mark or CGI is present, which is equivalent to the interaction term between each histone mark or CGI and RepD. Enrichment of hyperDMR is highlighted in red, depletion is highlighted in blue, and insignificant results shown in grey. Acronym: SE, *standard error*.

According to the results of the regression model, CGI, H3K27me3, and H3K4me1 are independently associated with increased chances of hyperDMR, on top of the background enrichment associated with late replication. On the other hand, H3K9me3, which is associated with heterochromatin, is associated with a decreased chance of hyperDMR.

H3K27me3 is one of the best known repressive histone marks, and is one of the characteristic marks for bivalent promoters. It can be stably transmitted across cell replication [295, 296]. After DNA replication, epigenetic marks including H3K27me3 are diluted, but the diluted H3K27me3 is still able to recruit Polycomb repressive complex (PRC) proteins, which replenishes H3K27me3 modifications to nearby histones. This self-propagation occurs perhaps in a sequence specific manner [296].

Both PRC and DNA methylation are repressive, but their interaction is complex and incompletely understood. Biochemically, PRC can interact with both DNA methyltransferases (DNMTs) [297] and Ten-Eleven Translocation (TET) [298] proteins, the latter being a key player in DNA demethylation. The biological

interplay between the two is best studied in the context of developmental genes, such as the HOX gene clusters, which are low in DNA methylation and high in PRC repression. An antagonistic relationship has been described in prostate cancer [190] and backed up with *in vitro* mechanistic insights using gene knockouts, where loss of *Eed*, a PRC protein leads to increased DNA methylation [299, 300], and loss of *Dnmt3b* leads to increased H3K27me3 [300].

Data analysis has shown that late replicating regions with H3K27me3 marks are enriched in hypermethylation, and literature suggests that when H3K27me3 is lost, DNA methylation increases. Taken together, it is reasonable to hypothesize that the replication stress may lead to a dilution of H3K27me3 marks across cell divisions, and ultimately leads to a reactive increase of DNA methylation. How this dilution occurs may be because of passive loss due to inadequate time to replenish histone marks, or secondary to the loss of DNA methylation resulting in a redistribution of H3K27me3 [301].

For CGI, it has been demonstrated that there may be an active mechanism related to local CpG density which protects CpGs from being methylated [302]. When disrupted, *de novo* DNA methylation would occur. Replication stress may also dilute or disturb these uncharacterized methylation-protecting factors, leading to CGI hypermethylation.

It is known that H3K4me structurally antagonizes DNMT3's binding to chromatin. An antagonistic relationship similar to that of H3K27me3 has been reported in a pan-cancer study [303], where decreased H3K4me1 is associated with increased DNA methylation. It is unclear why H3K4me3 is reported to have no significant independent association with hypermethylation in the regression model, but is probably due to the correlation between K4me1 and K4me3 marks.

Altogether, hypermethylation occurs in regions that are known to be or possibly actively excluded from DNA methylation. Late replication not only negatively impacts the inheritance of CpG methylation, but also impairs the stability of antagonistic factors which keep CGI unmethylated, leading to focal hypermethylation.

Bivalent promoter methylation has minor effect on transcription

Having known that focal hypermethylation may be secondary to the loss in H3K27me3, an immediate question is then whether bivalent promoter methylation has any effect on gene expression.

Loci to gene annotation is downloaded from GeneHancer v5.14 [304] using the double-elite definition, and mapped to promoter and bivalent promoter HMM annotations by direct overlap. Multi-mapped annotations are removed from analysis. Genes with conflicting DMR results (e.g. promoter being both hyper- and hypomethylated) are also removed. Moderated log2 fold change (L2FC) is obtained using the `ashr` algorithm [244] implemented in *DESeq2* for tumour versus non-tumour, with tumour content included as covariate. Tumour content of the RNA-seq data is obtained by RNA-seq deconvolution using BayesPrism [243] as previously described in [237].

For promoter elements, hypomethylation is associated with a small but significantly increased L2FC, whereas hypermethylation is associated with significantly decreased L2FC. On the other hand, hypomethylation and hypermethylation of bivalent promoters are not associated with significant change in L2FC, as shown in Fig. 4.7.

This result suggests that in bivalent promoters of cancer cells, DNA methylation may act as a redundant mechanism in addition to H3K27me3 to maintain a stable gene expression status. For promoters, the canonical relationship between promoter methylation and reduced transcription holds true.

Tumour may maintain methylation pattern at lineage-specific enhancers

The 15 HMM annotations are made up of 97 HMM states. Exploratory analysis is done to investigate the distribution of DMR across different HMM states by comparing the odds of being differentially methylated to the expected odds of randomly distributed DMR. Most sub-states within each annotation group have a consistent direction of enrichment, except for enhancers and weak enhancers,

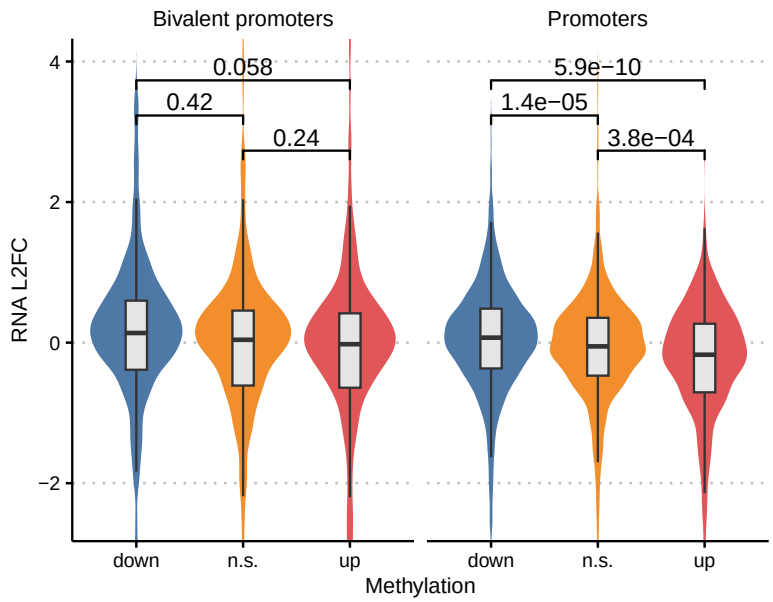


Fig. 4.7: Effect of DMR on gene expression in promoters and bivalent promoters. Moderated L2FC of tumour versus non-tumour is shown on the y-axis. The y-axis is limited between $[-2.5, 4]$ for visualization, but p value is calculated using t-test on the full range of data.

(Fig. 4.8), within which some states are strongly enriched for hyperDMR, whereas other states are strongly depleted for hyperDMR.

Within the enhancer annotations, EnhA7 to EnhA11 and EnhWk5 to EnhWk7 are specific for the haematopoietic lineage, and all of them are enriched for tumour hyperDMR. This may reflect that the non-tumour fraction of the biopsy is infiltrated with immune cells, which are probably hypomethylated at the lineage-specific enhancers, and therefore the tumour is relatively hypermethylated compared to non-tumour in these loci. On the other hand, EnhA12 to EnhA16 are annotated as digestive cell enhancers, which are depleted for hyperDMR in tumour. Since our samples are OAC, this suggests that tumours are able to retain lineage-specific enhancer epigenetic marks despite widespread aberration. Figure 4.9 shows the methylation of blood cells and digestive epithelium cells from [269] at the tumour-hypermethylated EnhA7 to EnhA11, and indeed the immune cells are less methylated, with myeloid cells being the least methylated. This suggests that the non-tumour compartment of the biopsies are indeed infiltrated with significant

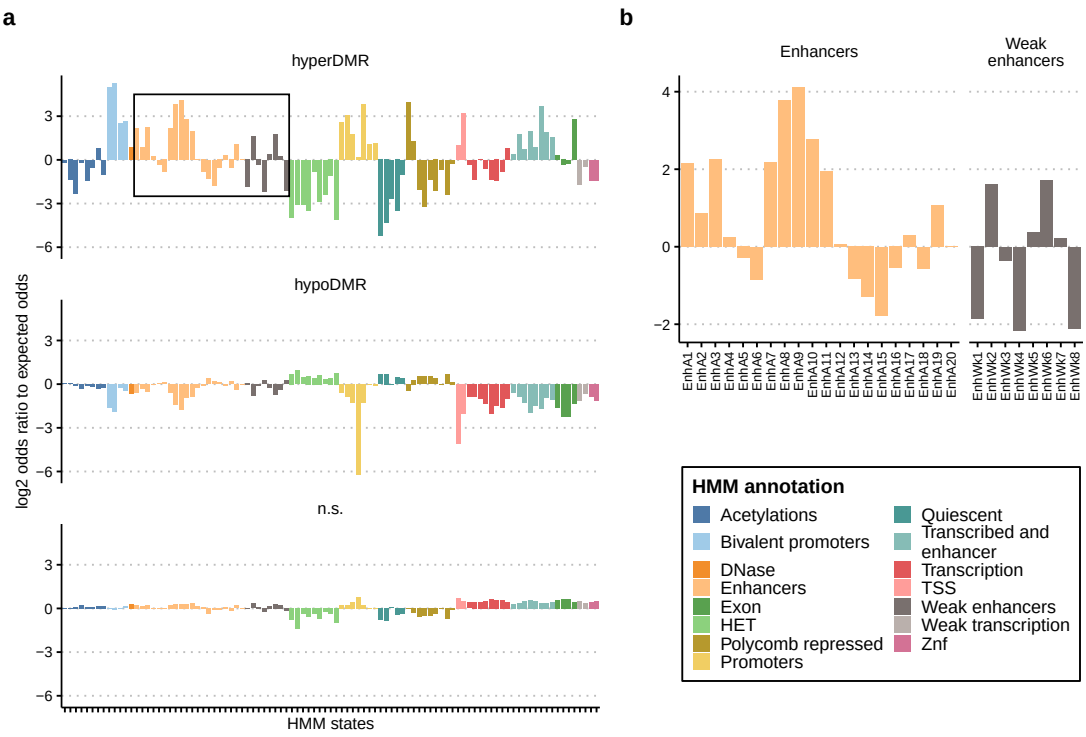


Fig. 4.8: Exploratory analysis of DMR across HMM states. (**a**) The y-axis shows the log2 odds ratio between the observed odds of DMR compared to the expected odds of a random DMR distribution. Most HMM annotation groups have consistent direction of odds ratio for each sub-state, except for enhancers and weak enhancers, as annotated using within black boxes. (**b**) Magnified view of enhancer states.

amount of immune cells, and performing DMR against adjacent normal references without considering tumour content may not be appropriate.

4.2.2 Investigation on specific genes of interest

Methylation panel for BO screening

Chettouh et al. has identified hypermethylation of TFPI2, TWIST1, ZNF345 and ZNF569 to be specific to BO compared to squamous oesophageal and gastric tissue [7] and can potentially be used in BO screening. However, the panel is defined using methylation array data, which may not capture the most characteristic differentially methylated regions. It is also uncertain whether the same sites remain methylated in OAC. Here I have validated that these sites are indeed hypermethylated in OAC, and have defined the DMR at single-base resolution as shown in Fig. C.9. In addition, I have identified additional hypermethylated loci in the neighbour of

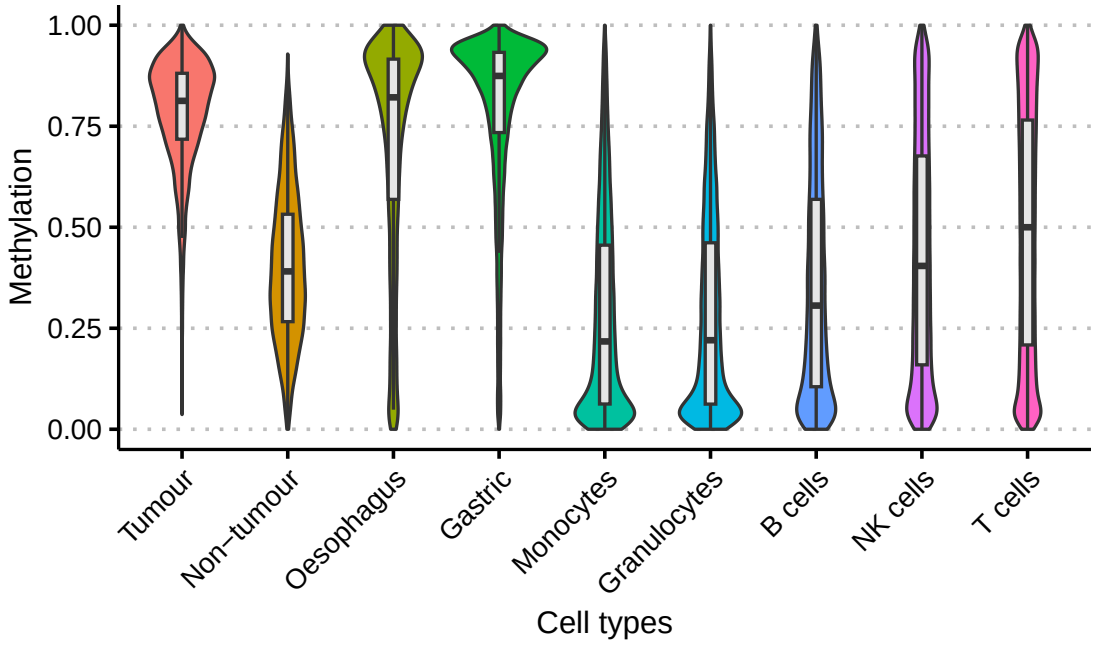


Fig. 4.9: Violin plot of methylation of normal cell types at haematopoietic lineage enhancers that are more methylated in tumour. Box plots are overlaid to show the median and interquartile range. Data is taken from Loyfer et al. [269], where the Gastric annotation includes antrum, body, and fundus epithelial cells, and immune cells annotations include their corresponding cell types from blood.

these genes. Interestingly, all 4 BO/OAC-hypermethylated sites are situated within broad hypomethylated domains. Considering broad hypomethylated domain may be a proxy for late replication, this finding is consistent with the notion that focal hypermethylation is also enriched in late replicating domains.

Role of CDKN2A

Literature review has identified CDKN2A to be an important player implicated in early development of BO and OAC [217, 225, 226, 305], as well as many other cancer types. It has been proposed that methylation silencing of CDKN2A (more technically, p16INK4a) may be a carcinogenic pathway.

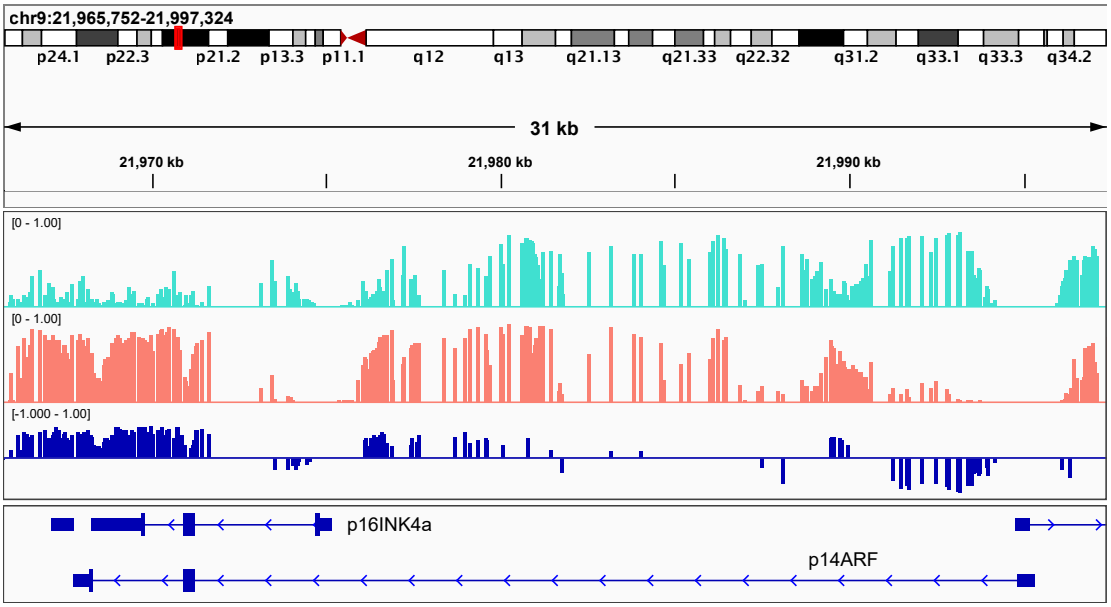


Fig. 4.10: CDKN2A. Top track represents the modelled non-tumour methylation (turquoise), 2ndtrack represents the tumour methylation (salmon), and 3rdtrack represents significant DMR (blue). DMR ranges from -1 to 1, with positive values being more methylated in tumour. The p16 and p14 transcripts are labelled respectively in the bottom window.

Figure 4.10 shows the modelled tumour and non-tumour methylation around CDKN2A in our OAC data. Two functional transcripts have been characterized for CDKN2A, the shorter one encodes p16INK4a, whereas the longer one encodes p14ARF (or p19ARF in mice), both being tumour suppressors [306]. Strong hyperDMR is found at the 3' end of CDKN2A and about 2 kb proximal to the p16INK4a transcription start site. In contrast, strong hypoDMR is found at about 2 kb distal to the p14ARF transcription start site. It is uncertain what overall functional consequences do the methylation changes have. RNA-seq analysis comparing tumour versus non-tumour does not show significant change in overall CDKN2A expression (L2FC = 0.60, fdr = 0.36). This may correspond to a report in 2005 showing the lack of relationship between hypermethylation and transcriptional silencing at this locus [307].

4.2.3 Independent validation of tumour-specific DMR shows pre-malignant onset

To validate tumour-specific DMRs discovered in the LUD2015-005 cohort, methylation array data are downloaded from the oesophageal cancer cohort (ESCA) of The Cancer Genome Atlas Research Network (TCGA) [222], the DOCTOR trial from Australia [283], and a related cohort from Krause et al. [231].

Due to differences in assay, a pre-filtering step is performed by comparing the methylation profile of normal oesophagi assayed by TAPS and Illumina HM450 methylation arrays. Due to the discrete nature of sequencing assays and high correlation between adjacent CpGs, window sizes of 0, 10, 20, 50, 100, 200, 500, and 1000 bp are constructed around each array probe for TAPS data to obtain smoothed methylation values. The smoothed values are then compared to array data of the same tissue type, and the window size with the highest overall Pearson's correlation is chosen for downstream analyses. A tumour-specific DMR on TAPS samples was rerun using the HM450 probe loci, and loci with fdr and DMR effect size in the top 0.1 quantile are chosen for replication.

The methylation of public datasets at the corresponding array probes are plotted against tumour content obtained from the original publications shown in Fig. 4.11. For the ESCA cohort, OSCC samples are also included. The methylation value shows a clear trend with tumour content in both OAC cohorts, and also shows a similar trend in OSCC, although the trend is less clear. This suggests that there are both tumour-specific methylation marks shared between OAC and OSCC, and also some marks that are specific to OAC.

Interestingly, although the methylation marks discovered in the LUD2015-005 trial are all from late-stage inoperable patients, the OAC-specific marks are present also in early-stage disease as shown in Fig. C.10. The Krause et al. dataset contains methylation data of BO, adjacent normal oesophagus and gastric tissue of tumour patients, nondysplastic oesophageal tissue from patients with GORD, and normal oesophagus from healthy individuals. It therefore provides an opportunity for us to further investigate the onset of the aberrant methylation marks in public dataset.

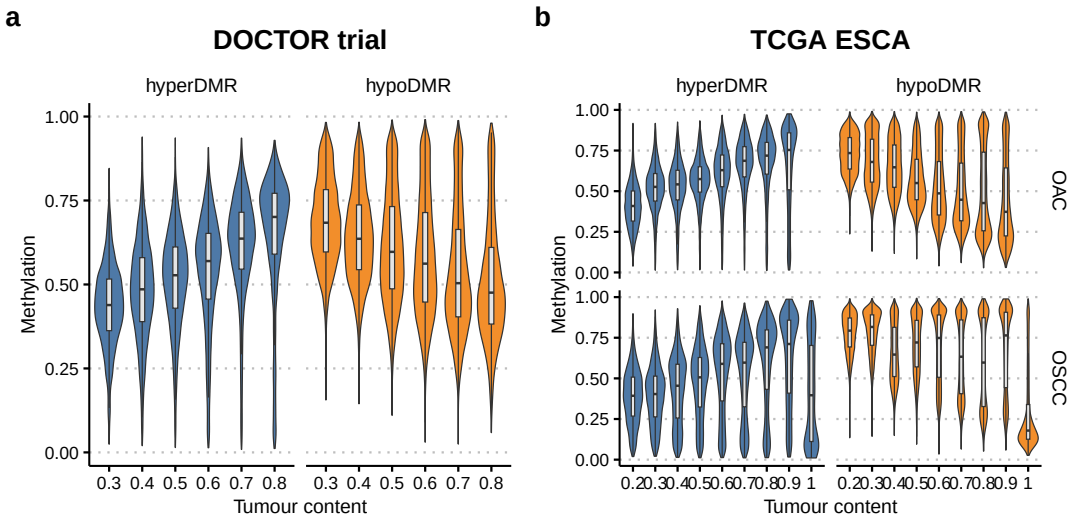


Fig. 4.11: Independent validation of OAC-specific methylation marks discovered in the LUD2015-005 trial. (**a**) Violin plot with methylation on the y-axis and tumour content in biopsy on the x-axis, binned into 0.1 intervals of samples from the DOCTOR trial. Box plot is overlaid to show median and interquartile range. Tumour-specific hypermethylation is coloured in blue, whereas tumour-specific hypomethylation is coloured in orange. The clear trend between methylation and tumour content suggests that these methylation marks are specific to tumour and not present in the tumour microenvironment. (**b**) Similar plot from the TCGA ESCA cohort. Top two panels show data from OAC samples, and bottom panels from OSCC. More data points deviate from the expected trend between methylation and tumour content.

As shown in Fig. 4.12, BO shows higher methylation than other tissues in OAC-specific hyperDMR, and lower methylation in OAC-specific hypoDMR. Oesophageal samples from GORD patients have similar methylation profile as adjacent normal and control oesophagi. This suggests that the onset of altered DNA methylation is between the inflammatory stage of GORD and the BO stage, potentially related, if not contributing, to the change in cellular plasticity from squamous to columnar cell type.

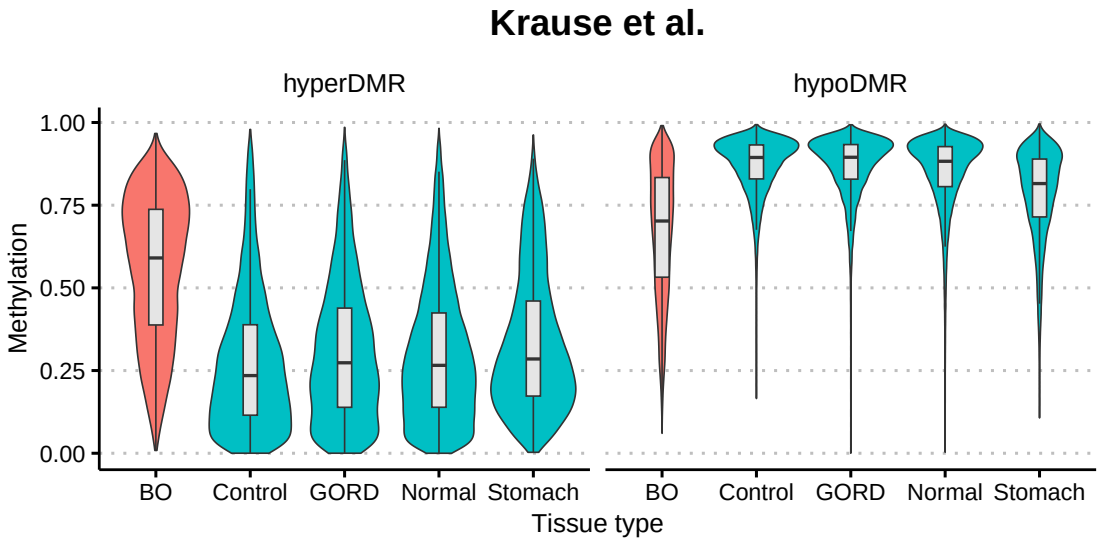


Fig. 4.12: OAC-specific methylation marks in BO. Violin plot with methylation on the y-axis and tissue type on the x-axis with overlaying box plot. BO is coloured in salmon and other tissues coloured in turquoise.

4.2.4 Detection of tumour-specific DMRs in cfDNA

As demonstrated in Fig. 4.2, broad hypomethylation is a common feature in OAC, and Fig. 4.12 suggested that this may be an early event in OAC development. On the other hand, Fig. 4.11b suggests that there is at least some degree of specificity for OAC compared to other cancer types. This led to the hypothesis that broad hypomethylation may be an interesting target for tumour detection or molecular characterization in cfDNA. A method for detecting broad methylation changes using binomial mixture modelling has been developed and detailed in Section 3.2.

Model performance using *in silico* diluted data

The model performance, in particular the false positive rate, using simulated data has been satisfactory as shown in Fig. 3.4. Yet, it is unknown how well it performs on real-life data.

To investigate this, tumour content in cfDNA is first obtained by running the *ichorCNA* [308] pipeline, a copy number caller based on sequencing coverage. 50kb window is used, with 0.4, 0.6, 0.8, 0.9, 0.95, 0.99, 0.995, and 0.999 used in the normal contamination setting. Subclonal estimation is turned off. Available

non-cancer control cfDNA samples ($n = 4$) are also included as panel-of-normal. Tumour purity estimates are taken from the default output without further quality control, and are shown in Fig. 4.13a.

Then, 4 pre-treatment (PreTx) OAC cfDNA samples with high and medium tumour content were diluted with control cfDNA in ratios of 50%, 25%, 12.5%, etc. The control cfDNA data used in the dilution is produced by mixing the `.bam` sequencing outputs of non-cancer cfDNA samples. The sequencing depths of the diluted samples are normalized to around 15x coverage, which allows some random sampling from the original data, which were on average sequenced to 30x.

A tumour-specific loci is defined as "positive" if there is more than 1 mixture components detected, and the largest methylation difference among the components is > 0.65. Then, a "tumour methylation detected" (TMD) metric is calculated by dividing the number of positive loci by the total number of loci tested for that sample. As shown in Fig. 4.13b, TMD follows the dilution curve, and the lowest dilution where the 4 samples can still be separated from control is 3.125%. This corresponds to a tumour content of around 0.2 to 1%.

Despite the potentially promising separation between control and cancer cfDNA samples, it should be noted that the same control samples were used to dilute the cancer samples. Although sampling variation has been accounted for by downsampling, the inter-individual biological variations in cfDNA methylation may be underestimated. Furthermore, non-cancer control has up to 0.13 TMD, demonstrating that the background noise in real life data can cause up to 13% false positive in the tested loci. In view of this, one may have to resort to defining TMD cutoffs based on population data, or refine the set of tumour-specific loci used to reduce false positive rates. Therefore, whether binomial mixture modelling can be used for differentiating cancer from non-cancer cfDNA samples at sufficient sensitivity and specificity requires further evaluation using an appropriate clinical cohort with adequate sample size.

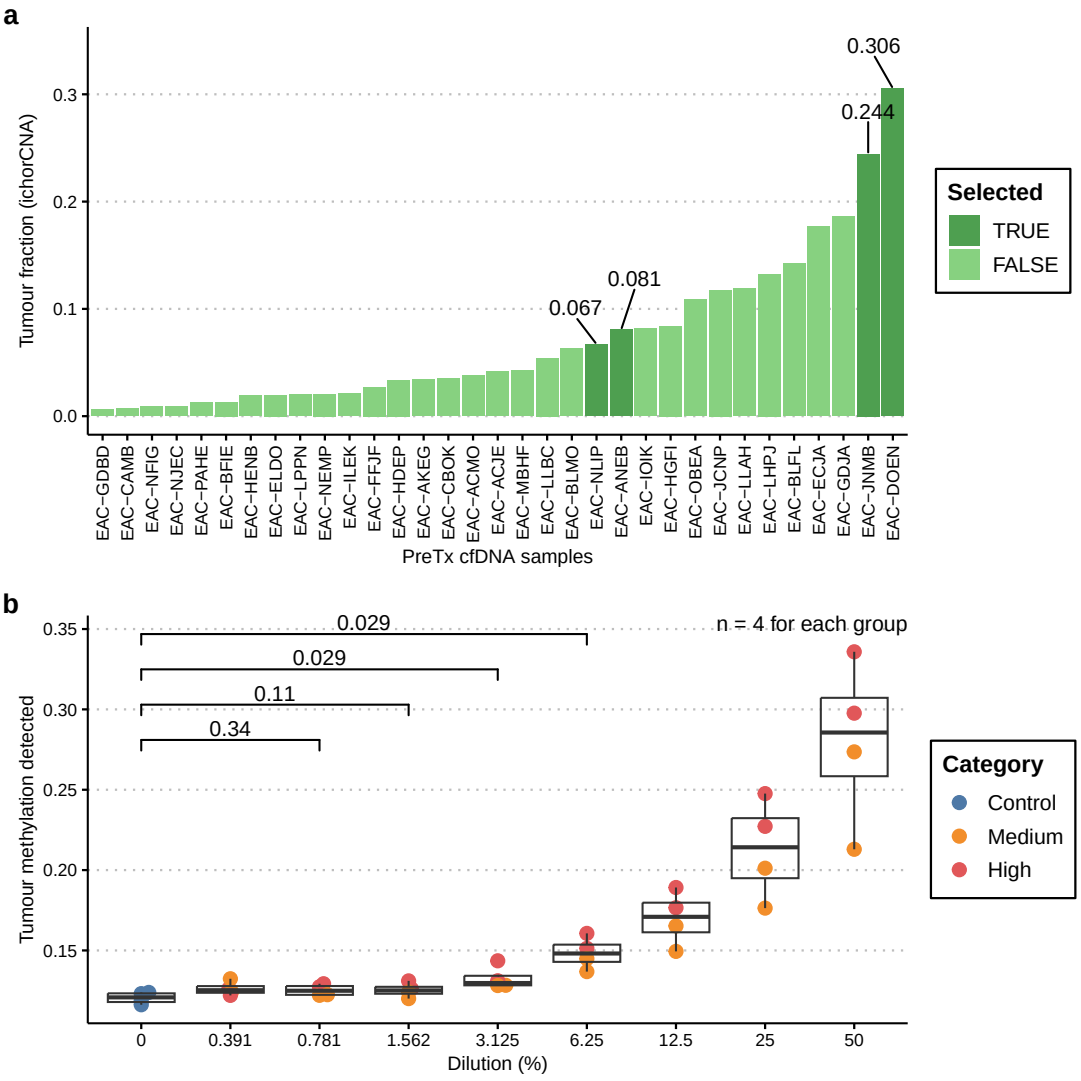


Fig. 4.13: Exploring the detection sensitivity of binomial mixture modelling. (**a**) Barchart showing tumour fraction in PreTx cfDNA samples from output of *ichorCNA*. 2 samples with high tumour content and 2 samples with medium tumour content were selected for *in silico* dilution, which are labelled in dark green bars. The selected tumour purity are also labelled accordingly. (**b**) Boxplot showing the TMD metric of various dilution ratio. Non-cancer controls are labelled with blue points, medium tumour fraction samples with orange, and high tumour fraction samples with red. Wilcoxon test is performed between non-cancer controls and selected dilution ratios, with p values shown above. The lowest dilution with a statistically significant difference is 3.125%.

Model performance compared with *ichorCNA*

The ability to detect tumour from low-depth sequencing is desirable in order to make clinical impact. Hence, the model performance of TMD is evaluated against *ichorCNA* at full sequencing depth (average 30x) and at downsampled depth of 1x coverage respectively. 1x coverage is chosen because at this sequencing depth, the cost of library preparation would become the dominant factor, and further reduction in sequencing cost would have minimal effect on the overall spending.

As mentioned in Section 3.2.4, tumour specific loci for 1x coverage is more lenient compared to that for 30x coverage due to inadequate read depth. Ideally the same set of loci should be used in 30x coverage for fair comparison, but computational time increases substantially if the lenient loci were used for 30x, and was therefore not performed.

The tumour content estimates using TMD and *ichorCNA* are calculated for all available cfDNA samples, with 4 non-cancer samples and 1 "long-term survivor recall" (LTSR) sample labelled as control. The LTSR sample is taken at approximately 3 years after the initial metastatic OAC diagnosis, and belongs to a patient with no residual tumour detected after treatment. Samples with <5% tumour content according to *ichorCNA* at 30x depth is plotted in Fig. 4.14. Figure 4.14a suggests that binomial mixture modelling demonstrates remarkable stability in the TMD estimates regardless of the sequencing depth and despite different sets of tumour specific loci being used. On the other hand, Fig. 4.14b shows considerably more variability in its tumour content estimate at low sequencing depth. Notably, *ichorCNA* detected much higher false positive tumour signal in one non-cancer control sample at 30x, whereas the false positives are relatively consistent across the control samples for TMD, which adds to the impression that *ichorCNA* may be less reliable at determining tumour content compared to TMD.

A comparison of TMD and *ichorCNA* estimates for all samples at 1x and 30x depth can be found in Fig. C.11.

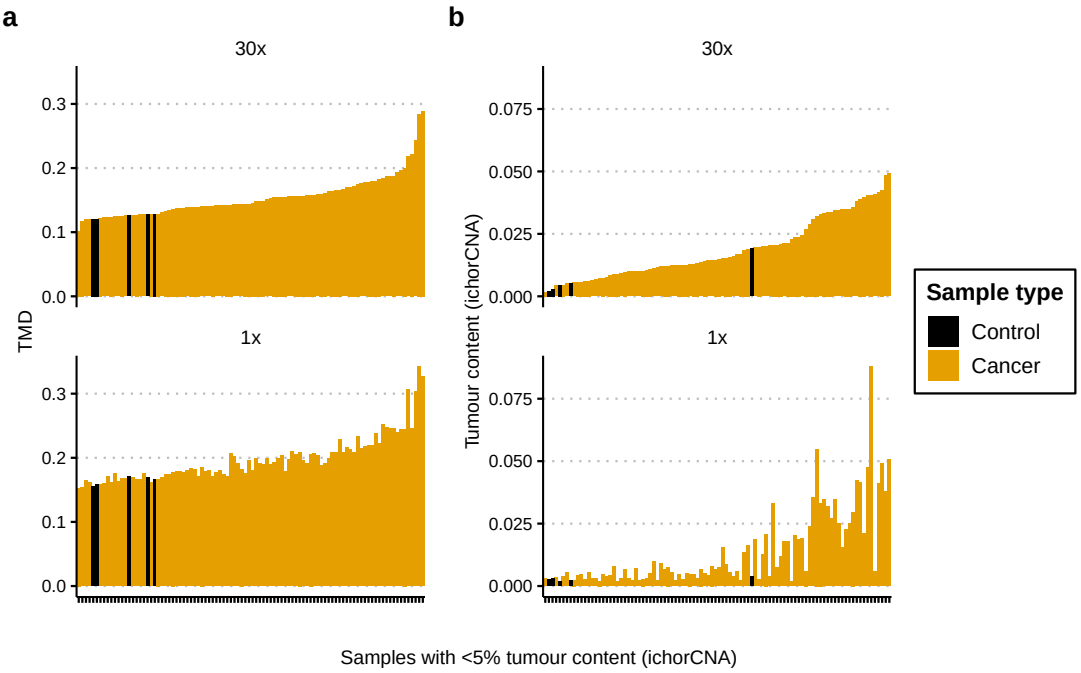


Fig. 4.14: Tumour content estimate at high and low coverage in samples with <5% tumour content according to *ichorCNA* at 30x depth. (**a**) Barchart showing TMD in cfDNA samples with low tumour content. Control samples, including non-cancer controls and long-term survivor recall, are represented in black and cancer samples in orange. Samples are ordered in the x-axis according to TMD at 30x depth. TMD estimate at 1x resembles that of 30x despite using a different set of tumour-specific loci. (**b**) Barchart showing *ichorCNA* tumour content estimates for the same cfDNA samples. Samples are ordered in the x-axis according to tumour content at 30x depth. The estimate varies substantially at 1x. Also, 1 non-cancer control sample has substantially higher tumour content estimate than the other controls.

Relationship with clinical outcome

It has been suggested that tumour content in cfDNA is a general prognostic marker for multiple cancer types [309–311]. In view of this, it may be interesting to investigate the relationship between cfDNA TMD and clinical benefit (CB). CB is defined as 12 months of progress-free survival (PFS).

As shown in Fig. 4.15a, patients with no clinical benefit (NCB) tend to have a higher TMD at baseline, although the difference is not statistically significant. Meanwhile, TMD at ICI-only timepoint shows a strong difference between the two groups with $p = 0.00047$.

Figure 4.15b shows the change in TMD with reference to PreTx baseline. It

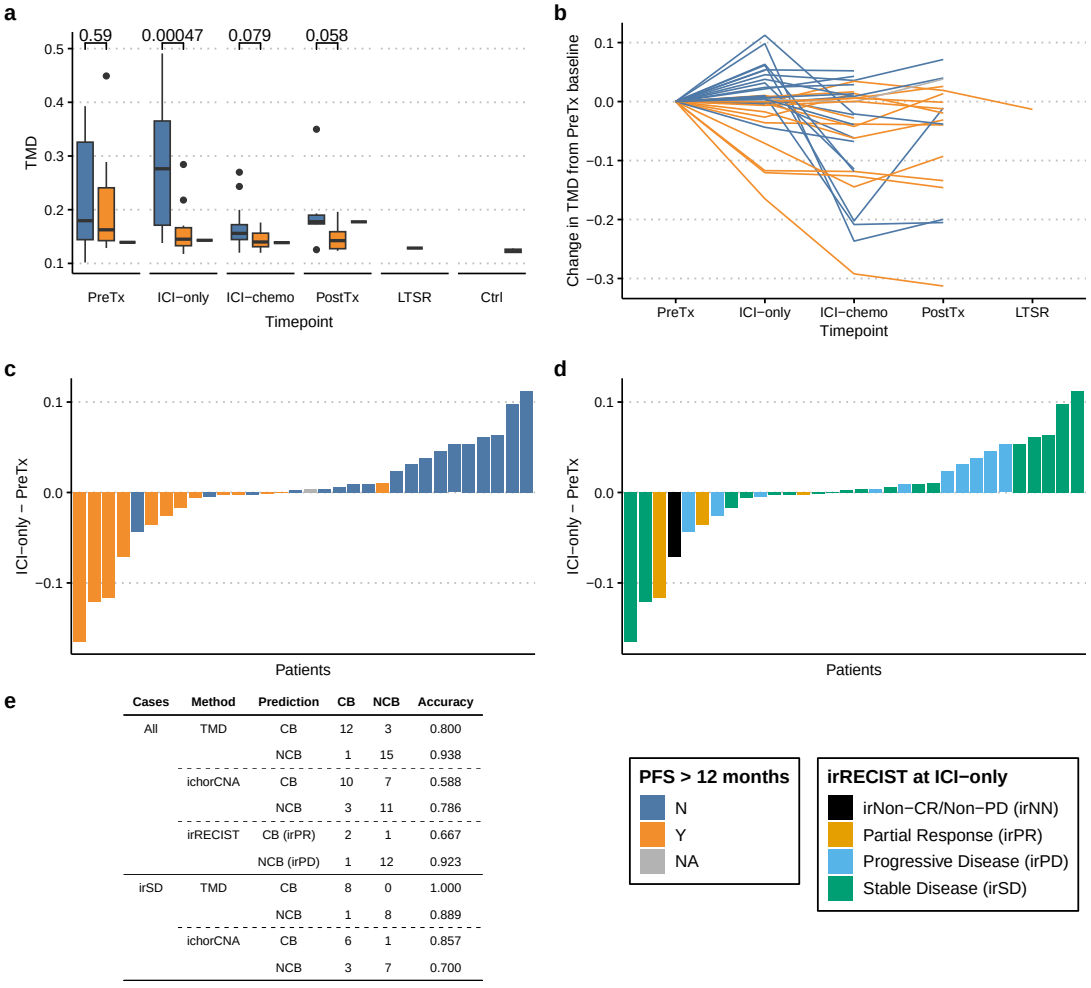
is notable that patients with NCB may have increased TMD at ICI-only, whereas those with CB have decreased TMD. At ICI-chemo, both CB and NCB patients have in general decreased TMD, which may rebound at PostTx timepoint, especially for NCB patients.

Figure 4.15c shows the change in TMD between ICI-only and PreTx for each patient, ordered from negative to positive difference. There is a stunning separation between CB and NCB patients, suggesting that dynamics in cfDNA TMD after 2 weeks of ICI may be prognostic of PFS at 12 months in late-stage OAC treated with combined immunochemotherapy. Figure 4.15d is the same plot, but coloured according to object response using the irRECIST criteria. Remarkably, change in TMD may be able to predict CB better than irRECIST, especially in cases where the patient is having radiologically stable disease (irSD).

A similar molecular response marker can be calculated using *ichorCNA* by subtracting the ICI-only tumour content by that of PreTx baseline, as shown in Fig. C.12. Suppose a positive change predicts NCB, and negative change predicts CB. To further assess the prognostic value of TMD, a contingency table is constructed in Fig. 4.15e, where the predicted clinical outcomes using different methods are compared with the actual outcome. TMD has the best accuracy at predicting both CB and NCB across all 3 methods. Additionally, when restricting to only cases of irSD, TMD only misclassified 1 out of 17 patients, whereas *ichorCNA* misclassified 4 out of 17.

It is noteworthy that 2 out of 3 misclassified NCB by TMD is classified as having progressive disease according to radiological evidence. This may suggest that TMD cannot replace clinical imaging, despite TMD having an overall better accuracy.

Lastly, the Kaplan Meier (KM) curves of overall survival (OS) and progression free survival (PFS) for TMD up versus down at ICI-only timepoint are shown in Fig. 4.16. Both median OS and PFS are not reached for patients with decreased TMD at ICI-only.



e

Cases	Method	Prediction	CB	NCB	Accuracy
All	TMD	CB	12	3	0.800
		NCB	1	15	0.938
	ichorCNA	CB	10	7	0.588
		NCB	3	11	0.786
	irRECIST	CB (irPR)	2	1	0.667
		NCB (irPD)	1	12	0.923
irSD	TMD	CB	8	0	1.000
		NCB	1	8	0.889
	ichorCNA	CB	6	1	0.857
		NCB	3	7	0.700

Fig. 4.15: TMD estimate at different timepoints in relationship with clinical benefit. Blue represents NCB, orange represents CB and grey represents N/A, which includes non-cancer samples and 1 patient whose progression status was ambiguous. (**a**) Boxplot of TMD at each timepoint. Wilcoxon test is performed between no CB and CB, and is significant only at the ICI-only timepoint. (**b**) Change in TMD with reference to PreTx baseline. Patients with NCB may have increased TMD at ICI-only, whereas those with CB have decreased TMD. At ICI-chemo, both CB and NCB patients have in general decreased TMD. At PostTx timepoint, there may be some rebound increase in TMD, especially for NCB patients. (**c, d**) Barplots of change in TMD at ICI-only at an individual level, ordered from negative to positive difference. The positions of samples are the same, but coloured according to CB in (**c**) and irRECIST in (**d**). TMD may predict CB in patients who were classified as having irSD. (**e**) Contingency table of the accuracy of predicting CB using the change in tumour content at ICI-only estimated with TMD and ichorCNA respectively. The prediction accuracy is calculated for all cases as well as for only cases with irSD.

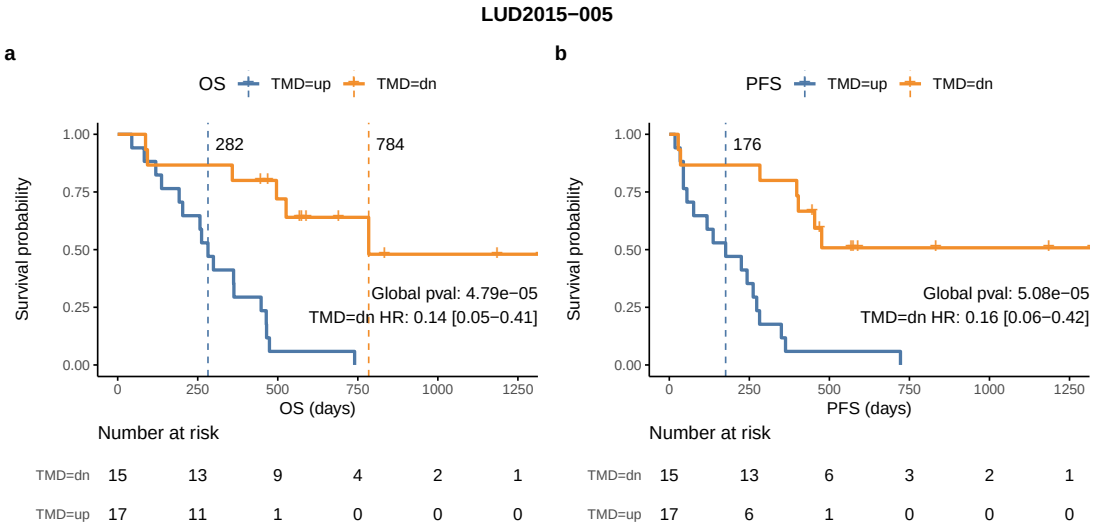


Fig. 4.16: KM curves for (**a**) OS and (**b**) PFS of patients with TMD up versus down at ICI-only timepoint. Median survival is represented by the dashed vertical line, with the exact number of days plotted next to the line. Median survival is not reached for TMD-down in both cases.

4.3 Summary and discussion

In this chapter, I have revealed novel insights into tumour biology by applying the statistical framework outlined in Section 3.1 to the tissue TAPS data from late-stage OAC patients in the LUD2015-005 trial.

Broad hypomethylation and focal hypermethylation has been found to be pan-OAC features, and both are more prevalent in late repD. These features have been validated using public DNA methylation array data, and may occur as early as in the stage of Barrett's oesophagus, a widely recognized premalignant stage of OAC.

While the mechanism of broad hypomethylation has been linked to passive loss of methylation during replication, why focal hypermethylation occurs in cancer remains elusive. By referencing public histone modification database, I identified strong independent associations between focal hypermethylation and features such as H3K27me3, H3K4me1, and CGI. This led to a testable hypothesis that these elements may contain factors that actively exclude DNA methylation, and replication may lead to a passive loss of these protective factors, ultimately leading to a reactive focal *de novo* methylation.

Promoter methylation has been proposed as one of the cancer-driving mechanism, I demonstrated using paired RNA-seq data from the trial that not all promoter methylation is equal. In fact, the DNA methylation status at bivalent promoters seem to have minimal effect on transcription, while DNA methylation level at regular promoters is inversely associated with transcript abundance. This may be because bivalent promoters are co-regulated by other epigenetic mechanisms such as repressive histone marks, and the redundancy in regulation maintains a stable level of transcription.

Apart from tumour biology, I have also demonstrated potential clinical applications of OAC-specific DMRs in cfDNA, using the framework outlined in Section 3.2. The use of broad DMRs enabled robust circulating tumour cfDNA detection below 1% with low-depth sequencing, and demonstrated superior sensitivity compared to imaging criteria in predicting clinical benefit after immunochemotherapy treatment

in late-stage OAC. This opens yet another avenue for using cfDNA methylation sequencing in cancer screening, as well as monitoring molecular response of solid tumours.

Intriguingly, the prognostic potential seems to be the strongest at the ICI-only timepoint, and not at chemotherapy-treated ones. This is interesting and unexpected, and suggests that first-line ICI may have a dominant effect in determining late-stage OAC survival in the trial, whereas the circulating OAC molecular response to chemotherapy defined in this work may not correlate with clinical outcome at 1-year after all. This might be due to multiple reasons, including chemotherapy interfering with tumour cfDNA release, chemotherapy affecting DNA methylation, and the eventual development of chemotherapy resistance.

It also remains a question of what does it mean when TMD increases after ICI, whether it represents increased cfDNA released from tumour cell death secondary to treatment, or increased tumour burden and dissemination, which may correspond to the recently described ICI-associated hyperprogression [312], possibly due to inflammation.

5

Inferences on biologically relevant methylation patterns

Contents

5.1 Overview

This chapter is focused on exploring the possible inferences that could be made from high throughput sequencing data based on limited sample number. In particular, given that the current understanding of methylation subtypes in OAC may be an artefact of tumour purity, can we still detect different OAC methylation subtypes? If yes, what are their respective molecular features and clinical implications? In

addition, what are the methylation signatures that are associated with good clinical outcomes?

As reviewed in [313], epigenetic silencing of tumour suppressor genes (TSGs) has been shown as one of the carcinogenic mechanisms. However, the evidence is limited to a few well-studied TSGs, and whether other aberrantly methylated TSGs can contribute to cancer remains questionable.

In my analyses presented in Chapter 4, both global hypomethylation and focal hypermethylation may be secondary to proliferative stress, instead of the other way round. Moreover, Fig. 4.7 suggests that changes in promoter methylation do not necessarily associate with a change in transcript abundance. This complicated two-way dynamics between methylation and cancer, on top of the presence of "silent epimutations", made the discovery of biologically relevant methylation changes rather difficult.

In addition to the above, as first-line ICI has only been approved for late-stage OAC recently in 2021 [11], there has not been any published omics profiling of ICI-treated OAC prior to our study, which poses great difficulty for us to replicate the outcome-related findings in our cohort.

Regardless, I hold the hypothesis, if not just wishful thinking, that at least some DMRs have big and consistent effect sizes across individuals, such as the case of being a key disease determinant or being downstream reactions to certain perturbations. By analysing the DMR together with orthogonal omics data such as WGS and RNA-seq, I hope my work can be a resource for future comers and inspire new research.

5.2 Results - tumour subtypes

5.2.1 Derivation of tumour methylation subtypes

It has been speculated in Chapter 4 that broad hypomethylation is a pan-OAC event, and the absence of hypomethylation may be explained by low tumour fraction. However, through data exploration on genome browser, it is observed that although hypomethylation is universal, the genomic location of its occurrence varies between samples, as illustrated in Fig. 5.1.

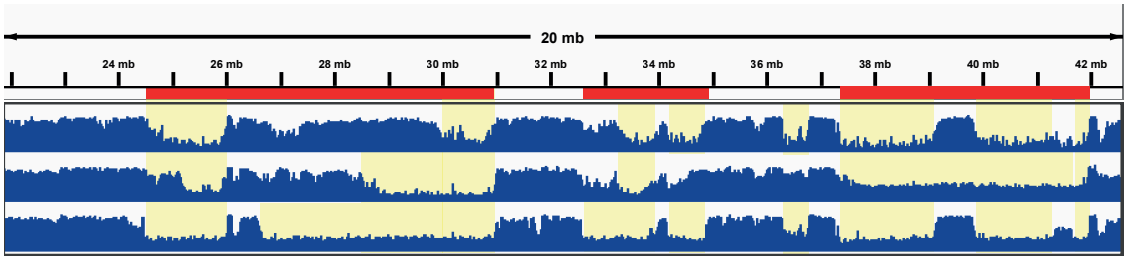


Fig. 5.1: Genome browser track showing the methylation pattern of 3 high tumour content samples over a 20 Mb window. Height of blue track represents DNA methylation, and the light yellow background indicates broad hypomethylation events, annotated visually. Genomic location of hypomethylation varies between samples.

It has been postulated that broad hypomethylation is caused by late replication. We therefore ask whether broad methylation patterns can classify OAC into molecular subtypes, which may suggest underlying differences in chromatin structure and replication domain (repD).

Data driven, tumour fraction and copy number aware subtyping was performed according to the frameworks described in Section 3.3. Both binomial test and mixture modelling approaches were applied to obtain the methylation status in 2 kb genome wide intervals constructed around CpG islands. The methylation status is coded as -1 for low methylation, 0 for average methylation, and 1 for high methylation. PCA was performed on the matrix of genome-wide methylation status for dimension reduction, and the transformed matrix was used for hierarchical clustering using the Ward D's method. The resulting dendrogram is shown in Fig. C.13. The optimal number of clusters is chosen to be 2 according to Calinski-Harabasz Index, and visualization with Uniform Manifold Approximation and Projection (UMAP) is also consistent with this choice (Fig. C.14). Cluster stability is assessed using the default options in `clusterboot` from *fpc* package in R, and Jaccard similarity is excellent upon bootstrapping (Fig. C.15).

Although different approaches for obtaining methylation status were used, both binomial test and mixture modelling resulted in the exact same clustering results. The cluster with lower global methylation is named "hypoT" (hypomethylated tumour), and the cluster with higher global methylation is named "hyperT". Note that this clustering method is based on trichotomized methylation states, which

do not take into account the possibility of coexistence of methylation subtypes. Assignment of a sample to a certain subtype cluster only suggest that the sample has a major representation of that subtype, and do not exclude the possibility of intra-tumoural heterogeneity. Unless otherwise specified, further clustering results will be based on the mixture modelling method.

A heatmap of the uncorrected methylation values is constructed in Fig. 5.2a for visualization. The top 2000 loci with the highest absolute weights in PC1 axis were chosen to generate the heatmap. Samples are grouped by the hierarchical clustering results, and ordered from low to high estimated tumour purity within each cluster. Non-tumour samples from oesophagus, stomach, and duodenum are also included in the heatmap.

Shown in Fig. 5.2b, the uncorrected methylation values decrease as tumour purity increases in hypoT, but not in hyperT, indicating that the selected loci are hypomethylated in hypoT subtypes but not hyperT. Compared to that of hypoT, the hyperT methylation landscape much more closely resembles that of non-tumour samples in these selected loci (Fig. 5.2c). This suggests that the biggest difference between hypoT and hyperT is probably broad hypomethylation in certain genomic regions.

The next immediate question is whether clustering by broad methylation pattern results in similar finding as previous published, i.e. CIMP-H, CIMP-L, etc. To answer this question, heatmaps are constructed using (1) CIMP-defining loci from Liu et al. and (2) CpG islands filtered similarly as that in Liu et al., which have < 0.3 average methylation in normal samples, and > 0.3 methylation in at least 1 tumour sample. Results are plotted in Fig. C.16. Intriguingly, both hypoT and hyperT demonstrate increased methylation at CpG islands when tumour purity increase. It is therefore unlikely that hypoT or hyperT corresponds to the published CIMP subtypes. Taken together with Fig. 3.7, which showed that CIMP-H samples may be misclassified as GEA-CIMP-L when tumour purity is decreased upon *in silico* dilution, a possible explanation is that CIMP subtypes in OAC could be an artefact from variation in tumour purity.

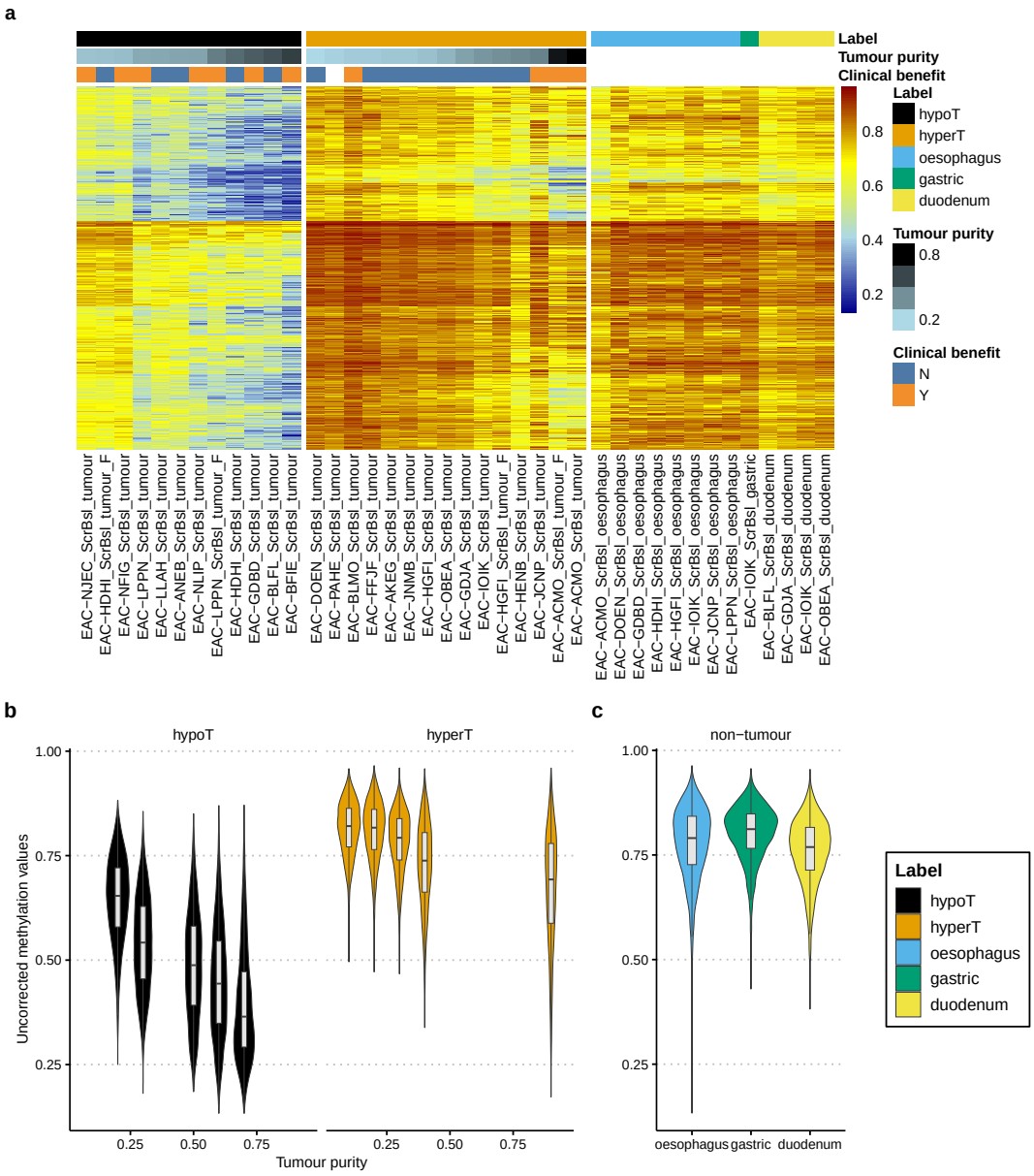


Fig. 5.2: Visualization of methylation values of tumour subtypes. (**a**) Heatmap of uncorrected methylation values at the top 2000 loci with the highest absolute weights in PC1. Samples are grouped by the tumour subtypes, as well as non-tumour controls. Tumour samples are ordered from low to high estimated tumour purity within each subtype. (**b,c**) Violin plots of uncorrected methylation values at the selected loci, of the two tumour subtypes in (**b**) and non-tumour controls in (**c**). A downward linear trend is observed for hypoT, but the trend is much weaker in hyperT.

5.2.2 DMR model performance

In addition to assessing the stability of subtype clusters by bootstrapping, the discovered subtypes are also assessed using AIC as described in Section 3.3.3. Briefly, DMR models are fitted using covariates of interest on top of tumour DNA fraction, such as hypoT and hyperT subtypes. Then, the model AICs are compared with that of a basic model with only tumour DNA fraction as covariate. As shown in Table 5.1, introducing subtype as covariate greatly reduces the AIC when building a genome-wide model based on HMM annotations, suggesting that the model performs better when tumour subtype is considered.

On the other hand, including CB in the DMR model results in an increased AIC, suggesting that the basic model is better. This does not mean there are no difference in methylation pattern between CB and NCB tumour samples, but rather including CB as covariate is unlikely to explain a lot of variance in the methylation data on a genome-wide scale. There may still be methylation loci that are functionally related to treatment outcome, although there would be a lack of appropriate public resource for the validation.

Genomic interval	DMR design	AIC	BIC
hmm	hyperT versus hypoT	402128.94	418244597.58
hmm	tumour versus stroma	28888921.69	279762114.04
hmm	CB versus NCB	36433473.25	453249630.03

Table 5.1: AIC and BIC with different DMR models, including tumour subtypes, CB, and only tumour fraction as covariate. Rows are ordered from top to bottom according to AIC. BIC is calculated but not interpreted due to reasons described in Section 3.3.3.

5.2.3 Molecular characteristics of tumour subtypes

Multiple analytical methods provided evidence that there are 2 major tumour subtypes, hypoT and hyperT, in our trial samples. Next, I will explore the molecular characteristics of the subtypes.

Methylation characteristics

The genome-wide methylation landscapes of the two subtypes are shown in Fig. 5.3. Consistent with Chapter 4, both subtypes demonstrate global hypomethylation in late repD. However, the hypomethylation in hypoT is more extreme, with almost no late repD remains methylated.

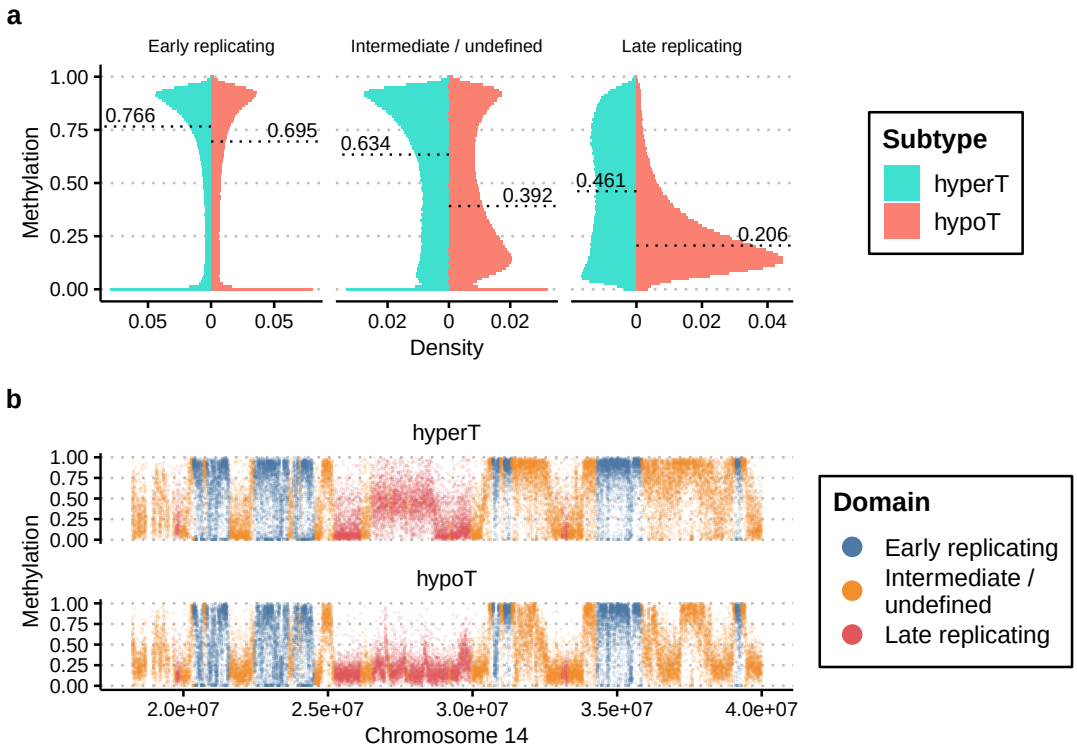


Fig. 5.3: Genome-wide methylation landscapes of tumour subtypes. (**a**) Mirrored histogram of methylation in different repD. HypoT has almost no methylation in late repD. (**b**) Methylation landscapes with sequence context. Colours represent different repD.

The difference in methylation patterns may either be acquired in the carcinogenesis process, or be related to the cell of origin. In fact, it has been suggested that cell of origin is a major factor in the molecular classification of cancers [314].

Using the top 2000 loci defined above as features, a hierarchical clustering is performed on the modelled tumour subtype methylation together with 207 normal cell types from Loyfer et al. [269]. The full heatmap is shown in Fig. C.20. Out of all cell types, hyperT subtype is closest to pancreatic acinar cells, whereas

hypoT is closest to a group of mesenchymal cells including fibroblast, smooth muscle, erythrocyte progenitors, and osteoblasts. Both groups of cells are unlikely to be cell-of-origin of OAC, suggesting that the difference in methylation pattern may be acquired in carcinogenesis.

Pancreatic acinar metaplasia (PAM) is a rare condition that can occur at the gastroesophageal junction, and can coexist with Barrett's oesophagus [315, 316]. PAM is found in the gastric mucosa as well, and is associated with *Helicobacter pylori* infection [317]. Yet, its association with cancer development is currently unknown. It is possible, though not entirely convincing, that hyperT has features of PAM, either suggestive of PAM being a pre-malignant process, or that this subtype shares some similarities with PAM as a response to inflammation.

Further investigation on the molecular evidence of PAM via lineage-specific methylation marks was not conclusive. Using markers from Loyfer et al. [269], neither subtype demonstrates significant hypomethylation in loci specific to pancreatic acinar cells (Fig. C.21). This may be confounded by CpG hypermethylation in tumour, or reflecting that the cell-of-origin of the subtypes is indeed not pancreatic acinar cells, or simply suggesting that hyperT is not related to PAM.

Mutational burden

Tumour mutational burden (TMB) has been widely established as one of the few independent predictors of immunotherapy outcome [318].

While Fig. 5.4a seems to suggest that hypoT has significantly higher median TMB, Fig. 5.4b suggests that this may be an effect of tumour purity. There is one outlier sample with extraordinarily high TMB, which was found to have microsatellite instability (MSI) in previous work [237].

It has been previously demonstrated that TMB has a non-linear relationship with tumour purity in WES, with only 64% of mutations still detectable at 80% normal contamination [21]. In our WGS data, the relationship seems to be rather linear, which might be attributable to the difference in sequencing depth.

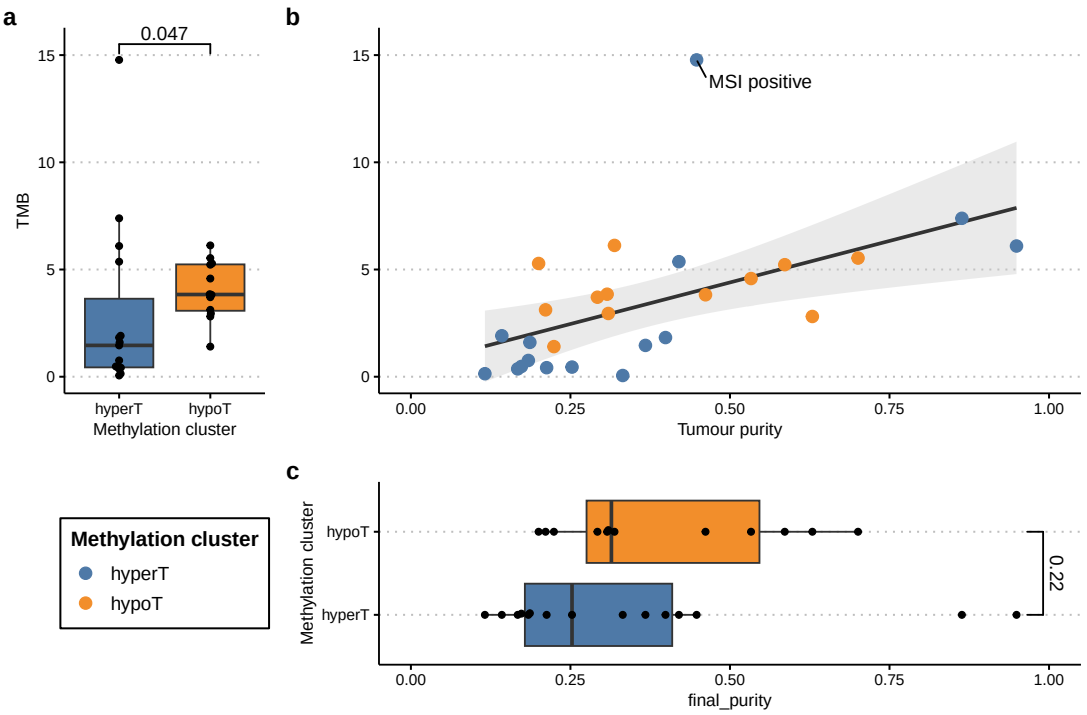


Fig. 5.4: TMB in subtypes, with blue representing hyperT and orange representing hypoT subtype. (**a**) Boxplot of TMB. HypoT has significantly higher TMB according to Wilcoxon test with $p = 0.047$. (**b**) Scatterplot of TMB against tumour purity. There is no visually discernable trend for the two subtypes. The outlier sample with extraordinarily high TMB has MSI. Black solid line represents a linear model fitted for TMB against tumour purity, with grey shadow as 95% confidence interval (CI). (**c**) Boxplot of tumour purity. HypoT has higher tumour purity, but the difference does not reach statistical significance.

A linear model of TMB is built against the subtypes with tumour purity as covariate with the MSI sample removed, as well as forcing a zero-intercept. Illustrated in Fig. C.17, there is no evidence for a difference in TMB between hypoT and hyperT.

Mutational signatures

Mutational signature analysis is carried out to investigate the possible aetiology of mutational processes in the tumour subtypes. *De novo* reconstruction of mutational signatures has been performed on a large cohort of OAC and Barrett's by Abbas et al. [230], and 14 single base substitution (SBS) signatures were identified to be the optimal configuration. Subsequently, these 14 signatures were used to evaluate

our cohort samples using *deconstructSigs* [251].

Unlike TMB, SBS signatures have not been reported to be affected by tumour purity to the best of my knowledge. A plot for mutational signature contributions against tumour purity can be found at Fig. C.18. Prior to FDR correction, significant associations between mutational signatures and tumour purity are found upon linear regression for SBS1, SBS8, and SBS17b. However, no significant hits remain after controlling for multiple testing. Intriguingly, SBS1, which is an age-related clock-like signature, is inversely correlated with purity.

Briefly, shown in Fig. 5.5, no statistically significant association is found between methylation subtype and SBS signatures. HypoT may have less SBS1, more SBS17a and SBS17b, but it is uncertain whether tumour purity is a confounder. A larger sample size and a formal appraisal on the effect of tumour purity on mutational signature identification are needed to thoroughly investigate this question.

The full heatmap of the 14 signatures in PreTx tumour samples can be found at Fig. C.19. SBS44 is low except in the sample with MSI.

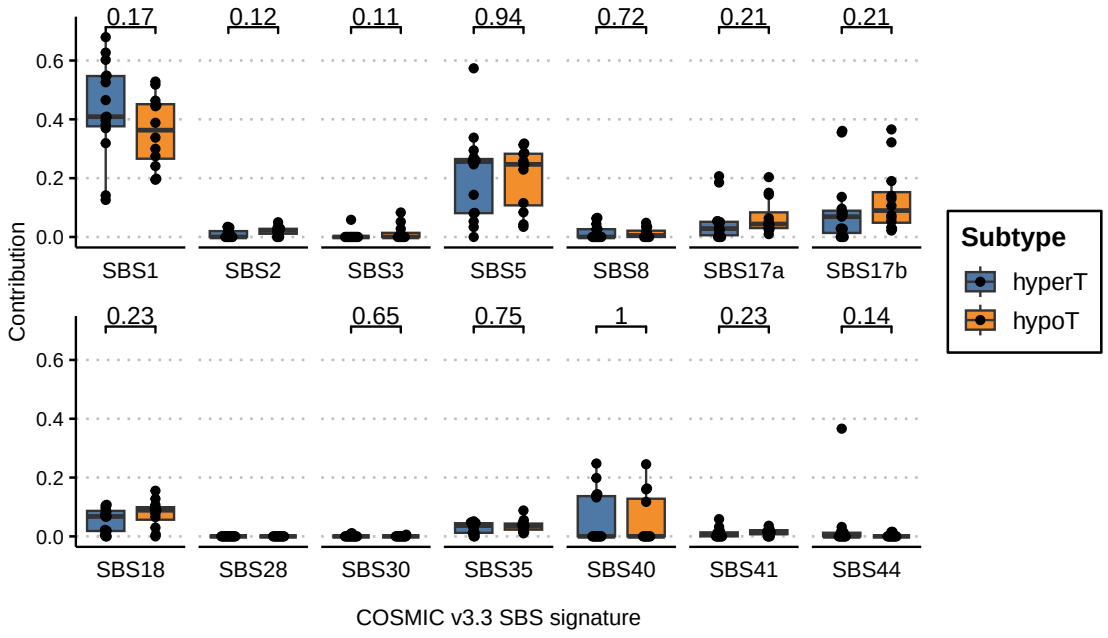


Fig. 5.5: Boxplot of mutational signatures in subtypes, with blue representing hyperT and orange representing hypoT subtype. Statistics is performed using Wilcoxon test. No statistically significant association is found between methylation subtype and SBS signatures.

Genome instability

A potential consequence of genome-wide hypomethylation is the reactivation of transposable elements, leading to genome instability.

Percent genome altered (PGA) is calculated as a measure of genome instability, which is defined as the percent of non-diploid genome or non-tetraploid genome for samples with whole genome duplication (WGD). WGD is defined according to Dentro et al. [255] based on a cutoff on the ratio between estimated ploidy and percent genome with loss of heterozygosity (LOH). The rationale was not well explained by the original authors, but one could reason that LOH has a higher chance to occur in a diploid state than a tetraploid state, because the former loses only 1 allele, whereas the latter needs to lose 2. This results in a high ploidy to LOH fraction ratio in WGD tumours.

Shown in Fig. 5.6, hypoT is potentially more likely to have WGD ($p = 0.087$), and indeed has higher PGA compared to hyperT ($p = 0.0017$). It is also noteworthy that close to half of the OAC samples in our cohort underwent WGD.

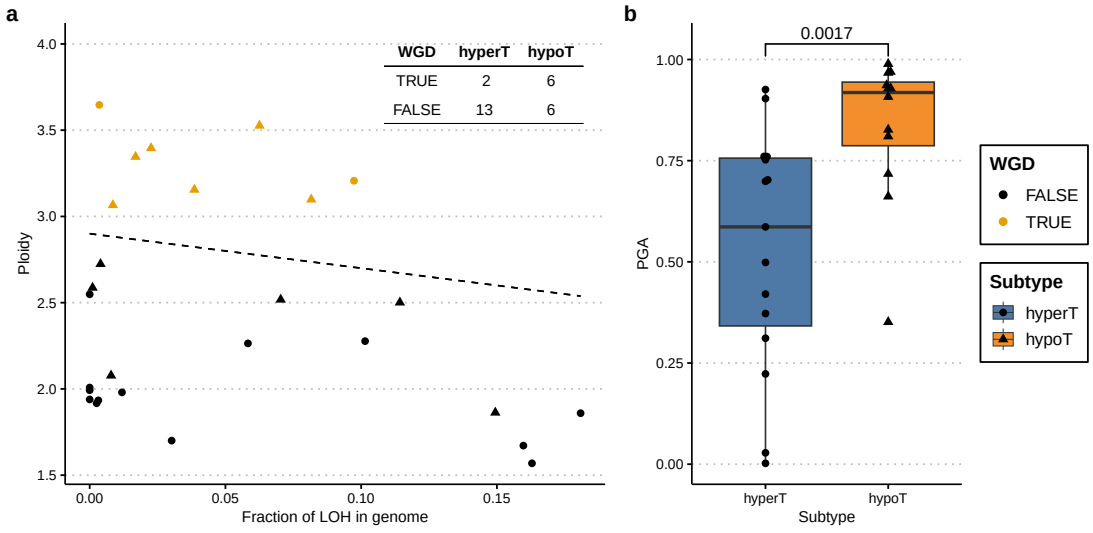


Fig. 5.6: Genome instability in tumour subtypes. (**a**) Ploidy of each sample plotted against LOH fraction. Colour of points represents WGD status, and shape of points represents subtype. Black line represents $y = 2.9 - 2x$, which is used for classifying WGD [255]. Contingency table at top right shows WGD may be more common in hypoT than hyperT (Fisher's test, $p = 0.087$). (**b**) Boxplot of PGA. Box colour and shape of points represents tumour subtypes, with hypoT having higher PGA (Wilcoxon test, $p = 0.0017$).

Top DEG hits and pathway analysis

Tumour purity adjusted differentially expressed gene (DEG) analysis of hyperT against hypoT is performed using *DESeq2* to characterize the transcriptomic differences between the subtypes, and moderated L2FC is obtained using the `ashr` algorithm. 25 genes has $fdr < 0.05$, and the variance stabilized gene counts are plotted against tumour purity, as visualized in Fig. 5.7.

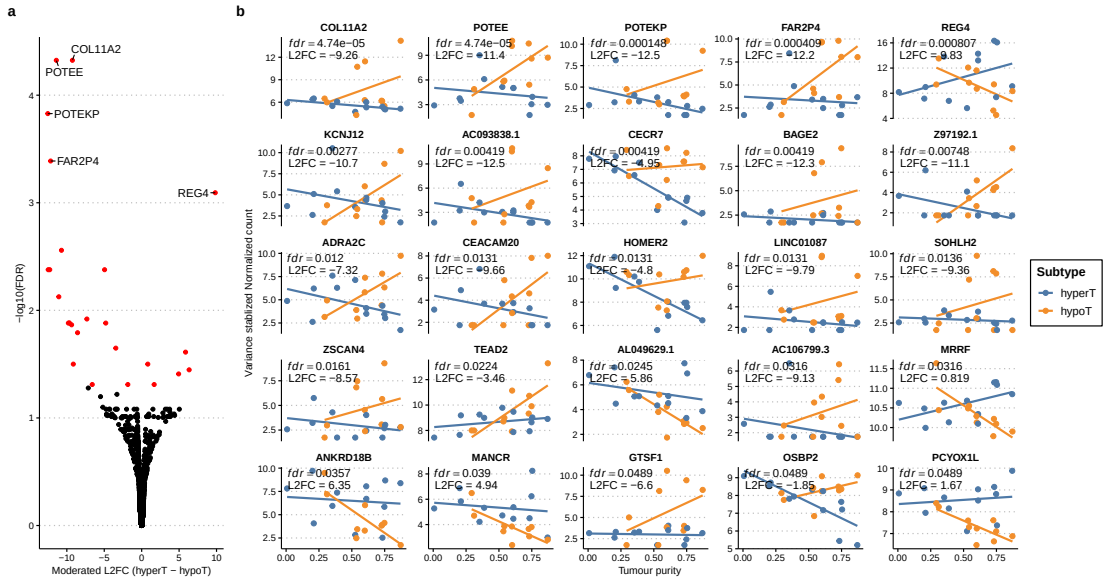


Fig. 5.7: (**a**) Volcano plot of DEG testing of hyperT against hypoT. Red points indicate $fdr < 0.05$. Positive L2FC suggests hyperT has higher transcript abundance, and negative L2FC suggests hypoT has higher transcript abundance. (**b**) Scatterplot of the variance stabilized gene counts against estimated tumour purity for the 25 significant DEGs comparing hyperT and hypoT. Colours represent subtypes. Facets are labelled with the gene name, and ordered from left to right and top to down according to p values. The fdr and L2FC for each gene are shown in the top left corner of each facet. A linear regression line is fitted for each subtype for visualization purpose only.

Genes that are more expressed in hypoT seem to be enriched in pseudogenes (POTEKP, FAR2P4, BAGE2) and testis-specific genes (SOHLH2, GTSF1, OSBP2). This may be explained by the extensively hypomethylated genome, leading to de-repression of methylation silenced genes.

Next, gene set enrichment analysis (GSEA) is performed using signed p value with FGSEA [245]. Hallmark gene set from MSigDB is used [246]. Shown in Fig. 5.8, hyperT is more enriched in proliferative pathways such as E2F, MYC. It is also more

enriched in epithelial mesenchymal transition (EMT) and G2M checkpoint pathways, suggestive of an overall more aggressive molecular phenotype. In addition, hyperT is enriched in response pathways to type I as well as type II interferons (IFN), although the latter did not pass the statistical threshold after FDR correction.

In contrast, hypoT may be more enriched in TGF-β signalling, heme metabolism, and myogenesis pathways, although none remains significant after FDR correction. TGF-β has been reported to be dysregulated in OAC [222], and have both tumour promoting and suppressing roles in gastroesophageal cancers [319]. On the other hand, heme metabolism and myogenesis are an interesting observation that corresponds to Fig. C.20, where hypoT is clustered closest to smooth muscle cells and erythrocyte progenitors.

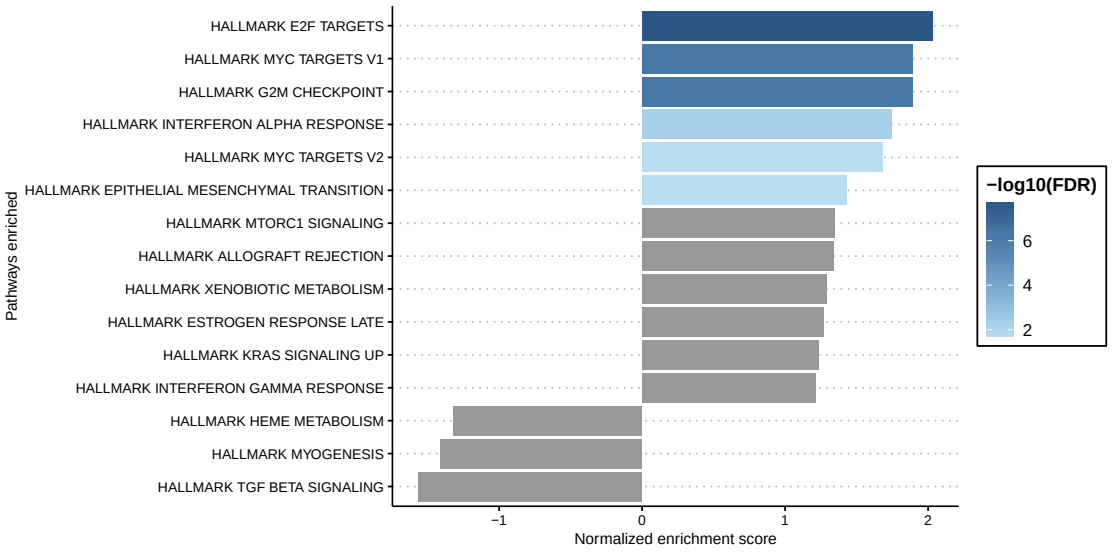


Fig. 5.8: Barchart of GSEA of transcriptome in hyperT versus hypoT, using the MSigDB hallmark gene set. Pathways with $p < 0.05$ are shown. Colour represent negative log10 fdr, and pathways with $fdr \geq 0.05$ are coloured in grey. Normalized enrichment score is taken directly from the output of `fgsea`. Positive score suggests the pathway is relatively enriched in hyperT, and negative score suggests the pathway is relatively enriched in hypoT.

Overlap between methylome and transcriptome

The local methylation landscapes around each of the 25 tumour subtype DEGs are manually inspected to look for potential transcriptional changes associated with aberrant DNA methylation. Long-range regulatory effects are not considered for this analysis.

The loci inspected can be roughly classified into 4 categories: (1) promoter hypoDMR and gene body hyperDMR associated with increased transcript abundance, (2) hypoDMR domain associated with increased transcription, (3) hypoDMR domain associated with decreased transcription, and (4) no clear evidence of association.

There are 4 tracks for all exemplar genome browsing plots in the following paragraphs. From top to bottom, they are:

1. Significant DMRs with $fdr < 0.01$ shown in red. Y-axis represents effect size from -1 to 1, where positive values indicate more methylation in hyperT.
2. Estimated hyperT methylation, shown in blue.
3. Estimated hypoT methylation, shown in orange.
4. Gene annotation.

Promoter hypomethylation and gene body methylation associated with increased transcription. Shown as follows is HOMER2, which is upregulated in hypoT. The hypomethylated promoter allows transcriptional activity, and the hypermethylated gene body is likely a consequence of active transcription through a 2-step recruitment of DNMT3B, which performs *de novo* DNA methylation [320].

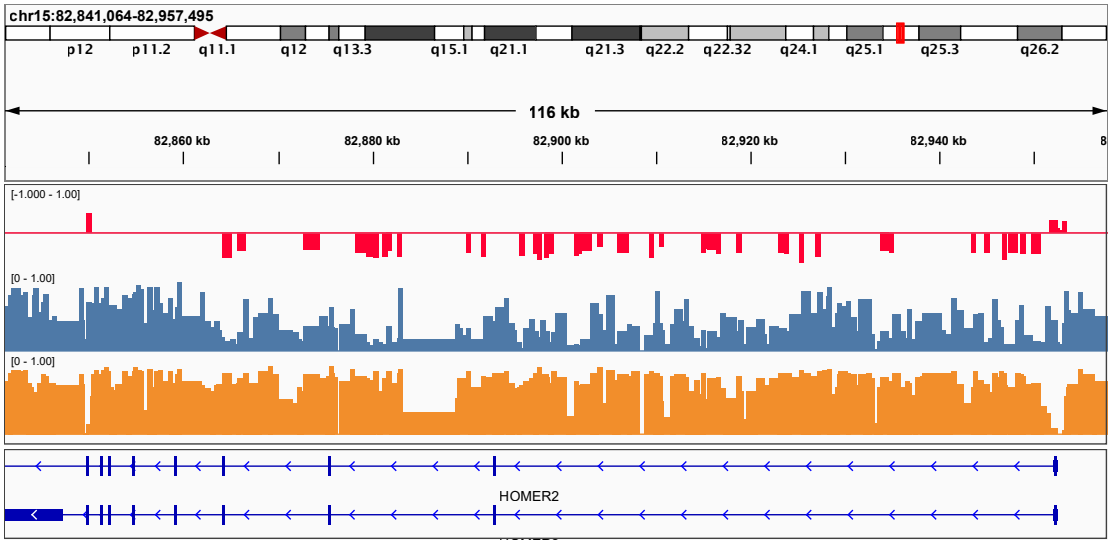


Fig. 5.9: HOMER2.

Another example is COL11A2, which is also upregulated in hypoT. Hypomethylated promoter and hypermethylated gene body is seen in hypoT.

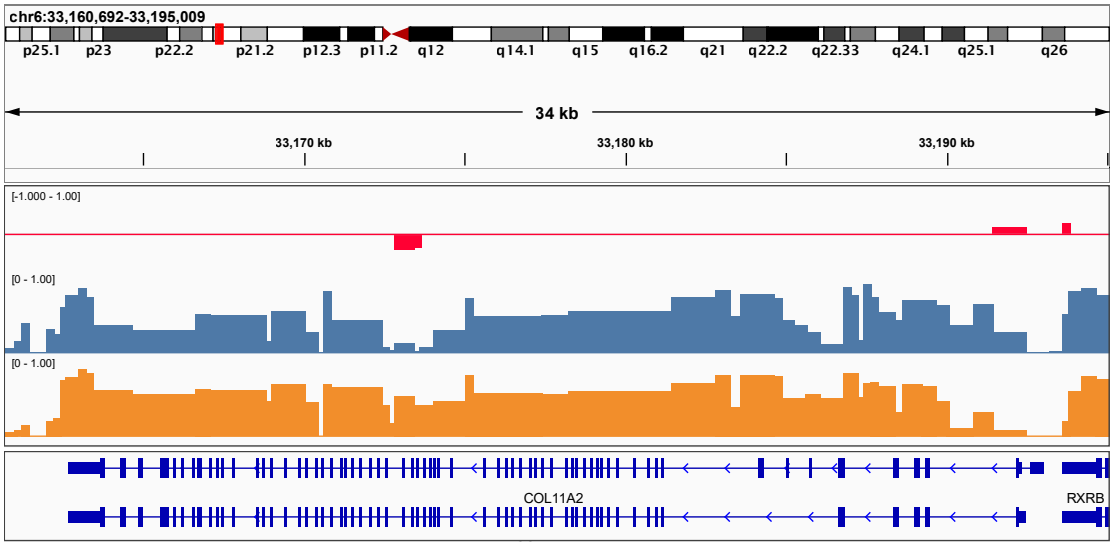


Fig. 5.10: COL11A2.

Hypomethylated domain associated with increased transcription. A 0.6 Mb window of chromosome 2 is shown here, with broad hypomethylation domains only in hypoT. 5 DEGs are found in this cluster, which are POTEE, FAR2P4, AC093838.1, POTEKP, and LINC01087 respectively. The genomic location of the 5 genes are marked in green above the DMR track. All 5 genes are upregulated in hypoT. Possible mechanism may be due to de-repression of methylation silenced genes. However, it is unexplained why the gene body of these regions are not methylated. One possibility is that these genes are not transcribed by RNA Pol II in hypoT but rather through other RNA polymerases that do not recruit histone methyltransferases and DNMT3A/B. This might suggest a different transcriptional machinery for genes inside silenced chromatin.

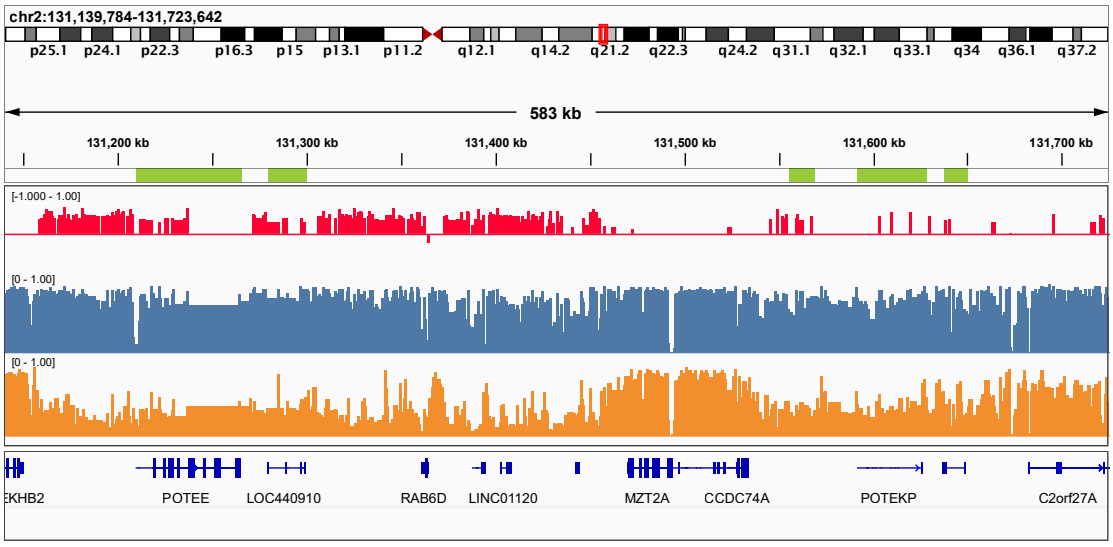


Fig. 5.11: Chromosome 2, 131.1 to 131.7 Mb, containing 5 DEGs.

Hypomethylated domain associated with reduced transcription. The entire gene body of ANKRD18B is hypomethylated. Figure C.22 provides a zoomed out view of the region, suggesting that ANKRD18B is at the edge of a hypomethylated domain. ANKRD18B is downregulated in hypoT, which is opposite case of the example above. Further inspection of the intercept of the DEG model suggests that ANKRD18B is expressed in the non-tumour compartment, and downregulated only in hypoT. A possible explanation is that this region is transcriptionally silenced by being "abducted" into a repressed chromatin domain, which subsequently lost DNA methylation due to late replication. In other words, the change in DNA methylation and transcription are both downstream to a chromatin modification event.

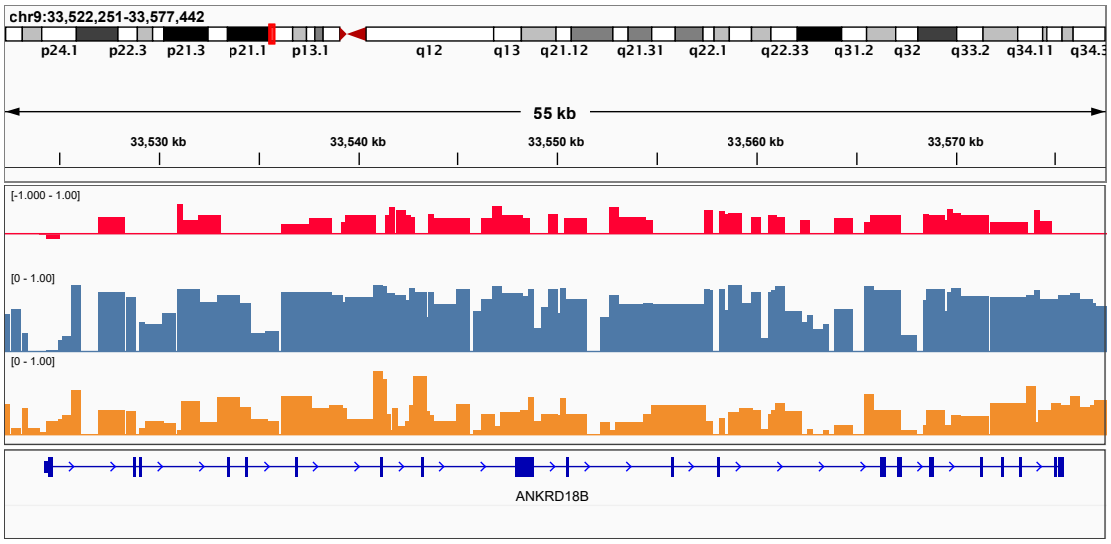


Fig. 5.12: ANKRD18B.

No clear evidence of association. Most DEGs have at least 1 associated DMR falling within the gene body and flanking regions. However, without knowing the exact regulatory regions of these genes, it is difficult to tell whether the DMRs are capable of causing the observed transcriptional changes. For example, KCNJ12 is highly upregulated in hypoT, but only 2 small DMR at the 3' untranslated region (UTR) is seen. Although gene regulation by distal element has been proposed in cancers [321], without further understanding of the existence of any distal promoter or enhancers or 3D chromatin landscape, it is difficult to infer the functional relevance of 3' UTR methylation in individual genes.

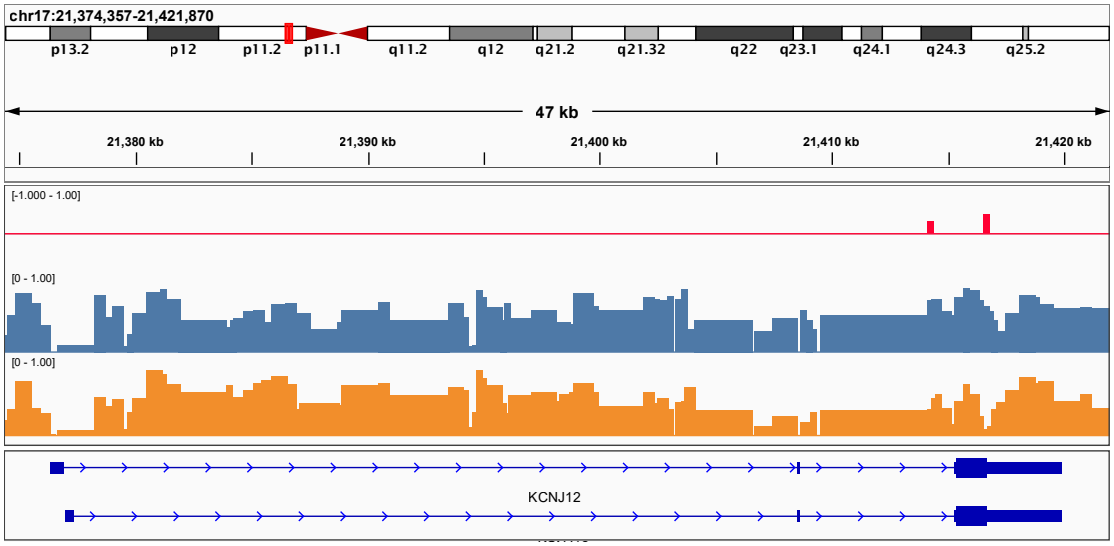


Fig. 5.13: KCNJ12.

5.2.4 Clinical characteristics of tumour subtypes

Figure 5.14 shows the demographics of the patients by tumour subtype. The distributions of age, sex, and disease stage are even among the subtypes for the given sample size.

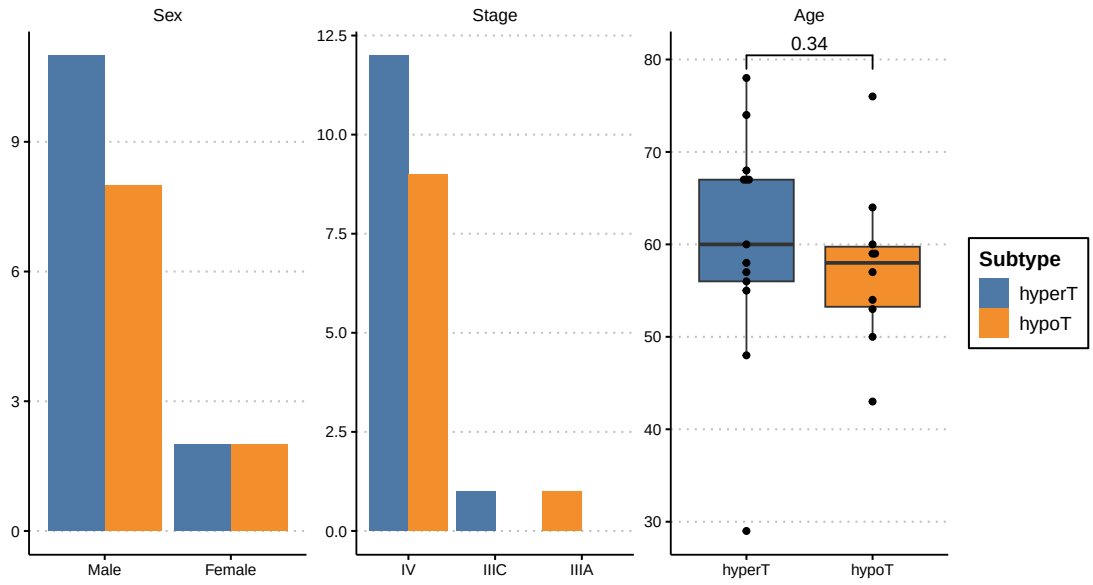


Fig. 5.14: Demographics by tumour subtype. Statistics is calculated for age using Wilcoxon test.

Kaplan Meier (KM) curves of overall survival (OS) and progression free survival (PFS) are shown in Fig. 5.15. Patients with hypoT-dominant tumours have a better trend of both OS and PFS, as well as an additional 4 to 5 months of median survival, although the differences are not significant.

5.2.5 Validation of tumour subtypes

Intuitively, to validate the presence of tumour subtypes discovered in the LUD2015-005 trial TAPS data, a set of subtype-specific loci needs to be defined, followed by testing the set of loci in public data. The markers should be able to classify the public tumour samples into hyperT and hypoT. Finally, these subtypes should have the similar molecular characteristics as described above, for example the transcriptional activities.

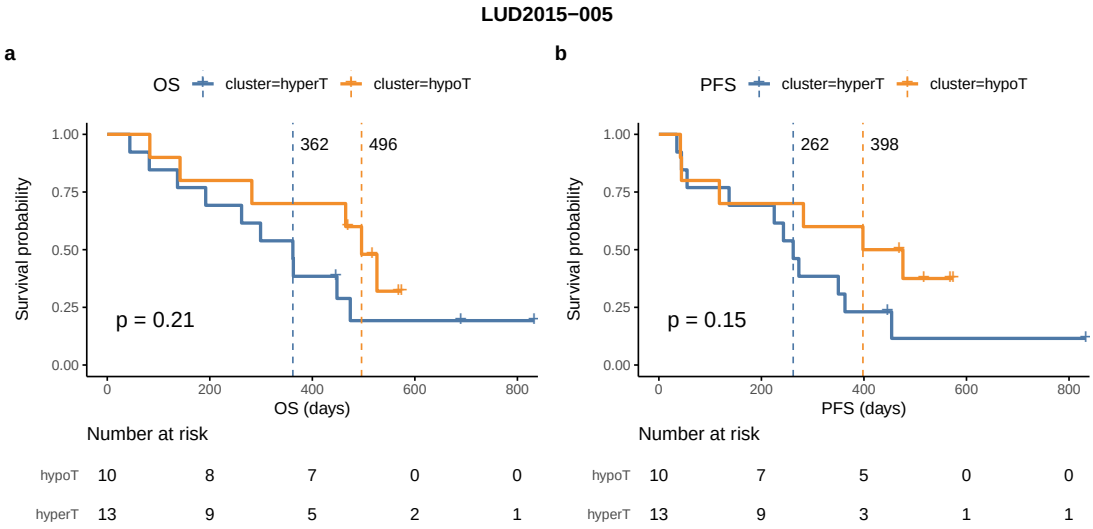


Fig. 5.15: KM curves for (**a**) OS and (**b**) PFS of the tumour methylation subtypes. Median survival is represented by the dashed vertical line, with the exact number of days plotted next to the line. HypoT has a better trend of OS and PFS, but the difference do not reach statistical significance.

Dataset and feature selection. Data from TCGA ESCA are used for the independent validation of tumour subtypes because of the easy availability of paired RNA sequencing data. Since the most striking difference between the subtypes are the broad hypomethylation pattern, these large-scale organized methylation changes are used as the subtype-specific loci instead. Similar to the procedure described in Section 3.2.4, CBS is used identify broad methylation changes, and a filtered is applied to select for loci with large difference. After filtering, 246 genomic regions of at least $1e5$ bp remain.

The next challenge is to detect the broad hypomethylation using array data, which do not have a good coverage outside CpG islands. Also, while the average methylation values inside broad hypomethylation domains are low, the variance remains high, meaning one could often detect highly methylated CpGs inside the domains. To tackle these problems, only CpGs in the "solo-WCGW" tetranucleotide context are chosen for validation. Solo-WCGW were defined in [209] as CpGs that are flanked by A or T nucleotides (therefore "W" according to IUPAC nucleotide code) on both sides, and have no other neighbouring CpGs within a +/- 35 bp

window. As shown by Zhou et al., these CpGs are especially prone to demethylation, and therefore may have a higher chance than other CpGs to be hypomethylated when inside a hypomethylated domain.

Classification. After feature selection, *InfiniumClust* is used to classify the OAC samples in the TCGA ESCA cohort, with consideration of tumour purity only and not copy number alterations. The distributions of methylation values in hypoT-specific loci are shown in Fig. 5.16. One of the cluster demonstrates a downward trend in methylation as tumour purity increases, whereas the other cluster remains methylated.

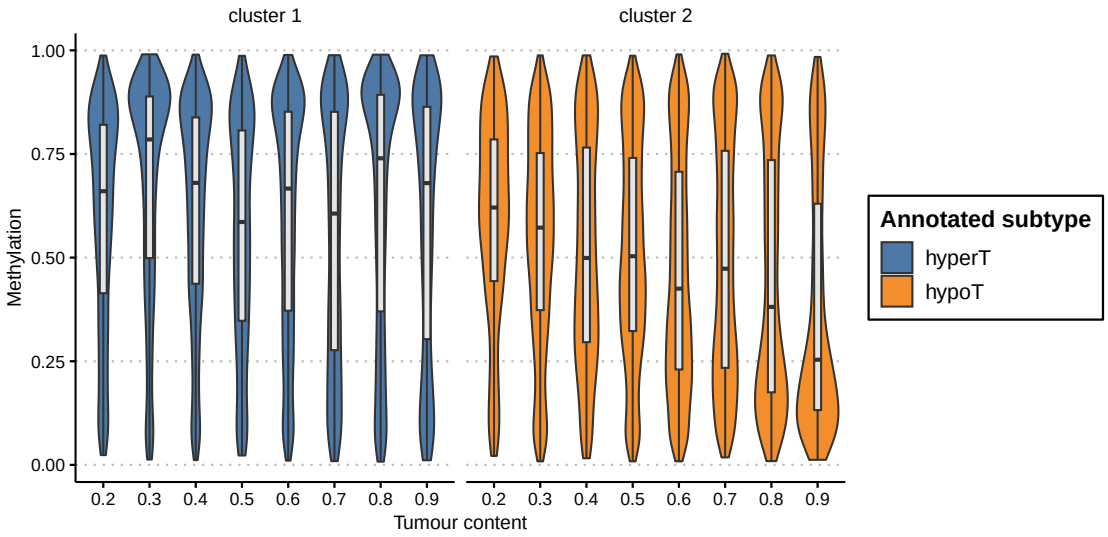


Fig. 5.16: Violin plots showing the distribution of methylation values in hypoT-specific loci against tumour purity in the two clusters, as classified by *InfiniumClust*. Cluster 1 is annotated as hyperT (blue), as it does not demonstrate a downward trend in methylation as tumour purity increases. Cluster 2 is annotated as hypoT (orange), as the methylation decreases when tumour purity increases.

Molecular characteristics. Next, TCGA ESCA transcriptome data is analysed in the same way as the LUD2015-005 trial data. Results of GSEA is plotted in Fig. 5.17, and are largely consistent with that shown in Fig. 5.8. The only conflicting exception is the response to IFN, which has an opposite enrichment in TCGA data, suggesting that this may be a false result in the initial discovery.

Otherwise, compared to hypoT, hyperT has an overall more aggressive picture. It is therefore concluded that the subtype-specific broad hypomethylation domains can be detected in array data, and the transcriptional characteristics are largely consistent between the subtypes.

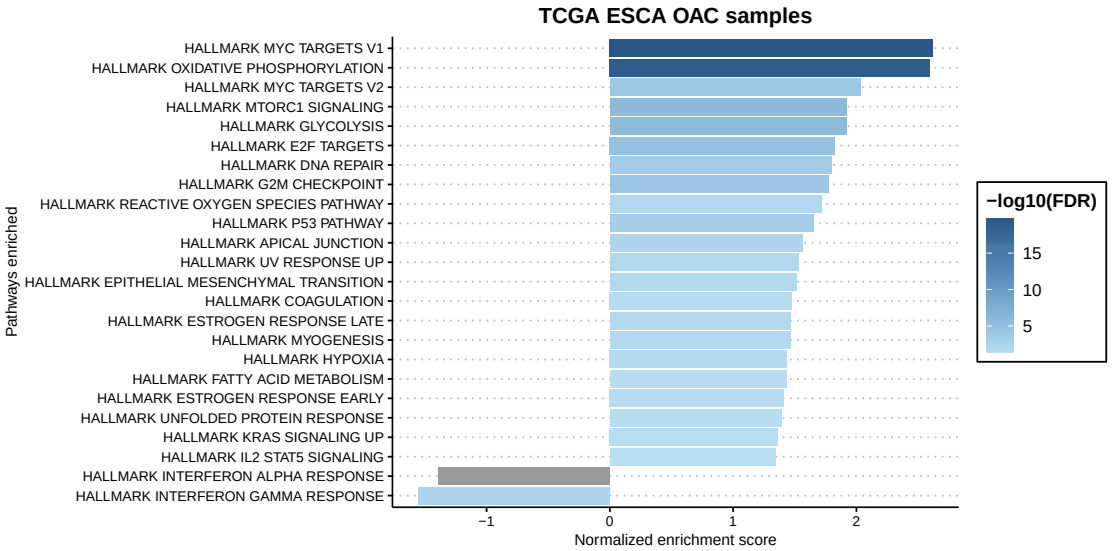


Fig. 5.17: Barchart of GSEA of transcriptome in hyperT versus hypoT in TCGA data, using the MSigDB hallmark gene set. As above, pathways with $p < 0.05$ are shown. Colour represent negative log10 fdr, and pathways with $fdr \geq 0.05$ are coloured in grey. Positive normalized enrichment score suggests the pathway is relatively enriched in hyperT, and negative score suggests the pathway is relatively enriched in hypoT.

Clinical outcome. Lastly, the KM curve for subtype-related OS in TCGA data is shown in Fig. 5.18, separated by stage I/II and stage III/IV disease. None of the survival statistics are significant. It is noteworthy however, that the median OS for patients with hyperT- and hypoT-dominant tumours are almost the same in late-stage disease, which is in contrast with that shown in Fig. 5.15.

Meanwhile, in early-stage disease of TCGA cohort, patients with hyperT-dominant tumours has much longer median OS than hypoT, but the survival probability continues downhill until it meets with the survival curve of hypoT. On the other hand, survival curve of hypoT has a long right-sided tail, suggesting that patients who survived the initial 1-2 years are unlikely to relapse.

One of the major differences between the two cohorts is that TCGA patients were not treated with ICI. If the observed trends are genuine, this suggests a possible survival benefit in treating hypoT-dominant OAC with ICI. A postulation of the mechanism is that hypoT expresses more pseudogenes and tesits-specific genes (Fig. 5.7b), and may harbour a higher number of neoantigens for immune cells to recognize. A larger DNA methylation dataset in OAC patients treated with combined ICI+CTX would be appreciated to further investigate this hypothesis.

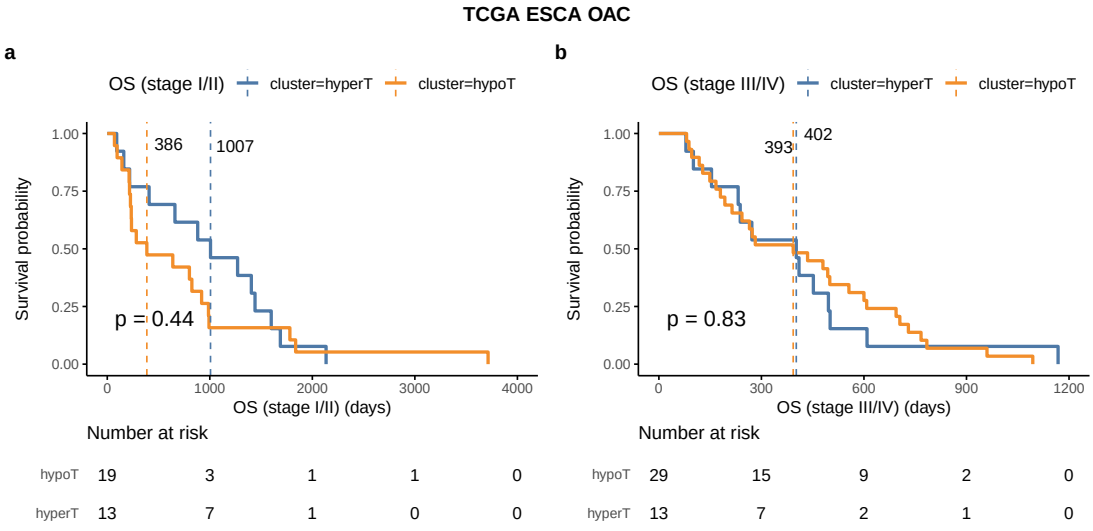


Fig. 5.18: KM curves for OS in (**a**) stage I/II and (**b**) III/IV disease in TCGA ESCA. HyperT has better median OS in early-stage disease, whereas both subtypes have similar median OS in late-stage disease.

5.3 Results - clinical benefit

In this section, "clinical benefit" (CB) is treated as the independent grouping variable, and analyses are performed to explore its association with other variables, such as DNA methylation and transcriptome data. The DMR model performance has been described above in Table 5.1.

5.3.1 Molecular characteristics associated with clinical benefit

Top DEG hits and pathway analysis

For CB versus NCB comparison in the tumour compartment, 24 DEGs has $fdr < 0.05$ (Fig. C.23). Of note, PD-L1 (CD274)expression is not found to be associated with clinical benefit in our cohort (moderated L2FC $= -0.176$, $fdr = 0.688$). Only 1 non-coding gene (AC002451.1) overlapped with significant DMR, which is not close to any known regulatory element. Unlike in the subtype comparison, there is a lack of overlap between DMR and DEG in the tumour compartment. It is therefore speculated that tumour methylation may not be a major determinant of ICI+CTX response in late-stage OAC. Instead, the focus of the analysis is switched to the tumour microenvironment, which may play a bigger role.

The tumour microenvironment is investigated by performing DEG on CB versus NCB in the non-tumour compartment. Instead of doing a model contrast to compare the slope coefficients of the regression model, a contrast of the intercept coefficients can be performed. There are 15 significant DEGs, one of which is IL10 (Fig. C.24), a cytokine produced primarily by monocytes. This is consistent with the finding by Carroll et al. [237] that pre-treatment tumour monocyte content is associated with better survival upon ICI+CTX treatment in the LUD2015-005 cohort.

GSEA is performed on CB versus NCB non-tumour compartment DEG using the cell type signature gene sets from MSigDB (C8). In summary, the microenvironment of CB samples are enriched in immune cell signatures, whereas NCB samples are enriched in gastroesophageal cell signatures. This suggests that CB samples are

associated with more immune infiltration at PreTx baseline. The full result of significant pathways enriched can be found at Fig. C.25.

Methylation characteristics.

Continuing from the hypothesis that CB samples are associated with more immune infiltration, DMR is performed for the non-tumour compartment, similar to how it was done in DEG analysis. Due to fewer number of significant hits, a more lenient $fdr < 0.05$ cutoff is chosen.

As mentioned in Chapter 4, the HMM annotation EnhA7 to EnhA11 are specific for haematopoietic lineage. Indeed, consistent with the hypothesis, 20 out of 21 DMRs that overlapped these enhancers are hypomethylated in CB microenvironment. The methylation of normal cells at these DMRs are plotted in Fig. 5.19, and suggest that haematopoietic cells are hypomethylated compared to gastroesophageal cells in these regions. It can therefore be proposed that these DMRs in CB microenvironment are contributed by immune infiltration.

Mutational burden

There is an apparent correlation between higher TMB and CB. However, as in the case above, the correlation becomes uncertain when tumour purity is considered (Fig. 5.20). Linear regression of TMB against tumour purity is performed using clinical outcome as covariate (Fig. C.27), and there is no evidence for a difference in TMB between CB and NCB.

Other molecular characteristics

The investigation for other molecular characteristics are largely negative findings. Plots are available for mutational signature Fig. C.28 and genome instability Fig. C.29.

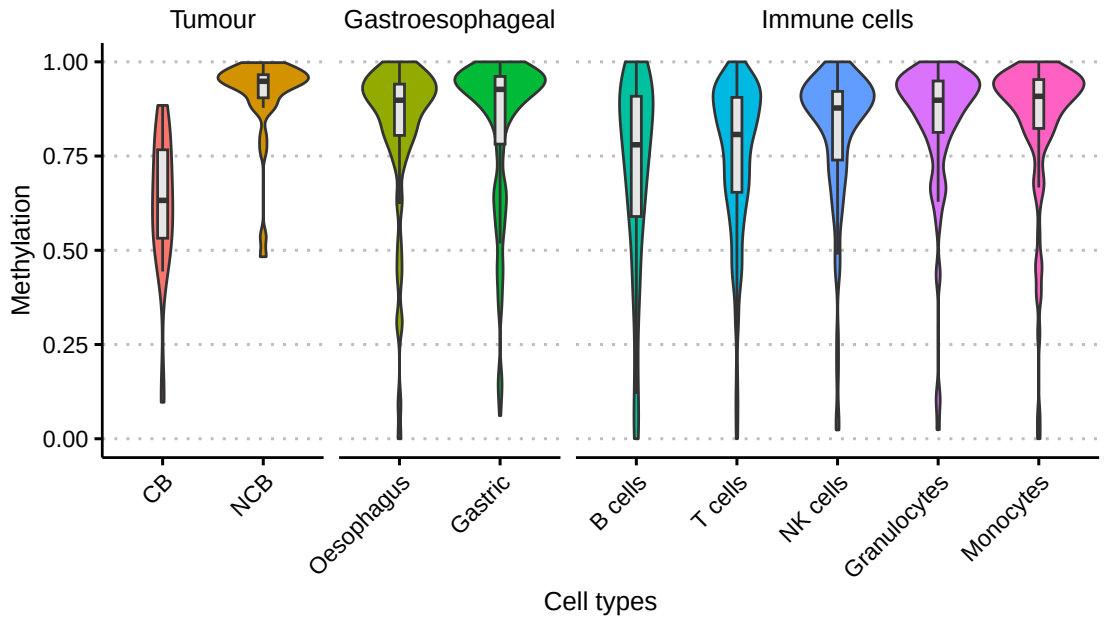


Fig. 5.19: Violin plot of methylation of normal cell types at haematopoietic lineage enhancers that are differentially methylated in the microenvironment of CB samples. Box plots are overlaid to show the median and interquartile range. Data is taken from Loyfer et al. [269] as previously described. Wilcoxon test is performed to compare median methylation between gastroesophageal and immune cells, with $p < 2.2e{-}16$ (not shown in plot).

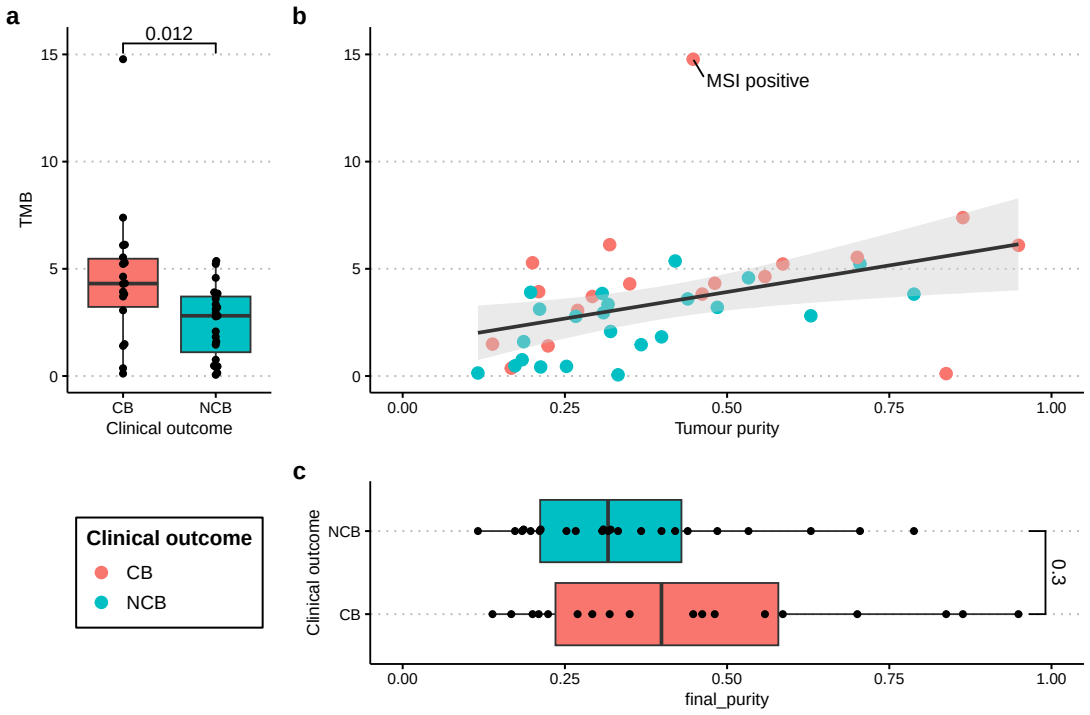


Fig. 5.20: TMB in tumour samples, with colour representing clinical outcome. (**a**) Boxplot of TMB. CB has significantly higher TMB according to Wilcoxon test with $p = 0.012$. (**b**) Scatterplot of TMB against tumour purity. There is no visually discernable trend for the different clinical outcomes. Black solid line represents a linear model fitted for TMB against tumour purity, with grey shadow as 95% CI. (**c**) Boxplot of tumour purity. CB samples have higher tumour purity, but the difference does not reach statistical significance.

5.4 Summary and discussion

In this chapter, I tackled the computational challenge of tumour subtyping in samples with variable tumour purity using bulk methylation sequencing data, and showed that OAC can be classified into 2 methylation subtypes, hyperT and hypoT. The subtypes are subsequently validated in TCGA ESCA cohort.

The most apparent molecular features of hypoT are the near-complete extent of hypomethylation in all late replicating domains, as well as the severe genome instability as evidenced by the high rates of whole genome duplication and aneuploidy events. Linking the stories together, it is possible that hypoT tumours are subjected to mitotic stress, which caused loss of DNA methylation in late replicating regions. The broad hypomethylation subsequently led to transposon reactivation, and ultimately genomic instability. HypoT tumours also have a higher expression of pseudogenes and testis-specific genes compared to hyperT.

It is rather counterintuitive, therefore, that the pathway analysis of transcriptome data suggests that hyperT is more enriched in the proliferation related pathways at pre-treatment baseline, such as E2F and MYC targets. Literature has suggested that global hypomethylation is proportional to the number of replication cycle, and if hyperT is indeed more proliferative, it should be more hypomethylated. Therefore, additional factors must play an important role in the replication timing related loss of DNA methylation. Possible explanations are that either hyperT is more efficient at maintaining DNA methylation than hypoT, or that the number of mitotic cycle only determines the "depth" of hypomethylation, but not "breadth", which is the key difference between hypoT and hyperT tumours.

Regarding mutational signature, there are no significant findings that stands out. HyperT has a higher SBS1 signature, which is an age-related clock-like signature. This could simply be due to a higher median age of patients with hyperT tumours than that of hypoT. However, many uncertainties lie in this result, and neither differences in SBS1 nor age are statistically significant.

The clinical response to combined immunochemotherapy also seems to differ between the two groups. Although the statistical analyses are again not significant

within the LUD2015-005 trial, the median overall survival and progression-free survival of patients with hypoT-dominant tumours are longer than those of hyperT-dominant by 4 to 5 months. This is in contrast with what has been observed in late-stage OAC of TCGA cohort, which is not an immunotherapy trial, where the median overall survival of the two tumour subtypes only differs by 1 week.

I also demonstrated 4 modes of how methylation may be related to transcript abundance, and showed that the transcriptional differences between the tumour subtypes are associated with corresponding methylation changes. There are cases where the gene may be de-repressed due to loss of methylation silencing, but also cases where both transcriptional and methylation changes may be secondary to a third factor, such as modifications in chromatin structure.

Last but not least, I raised a potentially important negative finding, which is the lack of association between tumour mutational burden and clinical response when tumour purity is taken into account. Apart from cases with microsatellite instability, tumour mutational burden demonstrates a linear relationship with tumour purity in our 60x coverage WGS. It is therefore crucial for future studies to account for tumour purity before studying the effect of mutational burden, which has major implications when considering mutational burden as a biomarker.

6 Discussion

Contents

6.1 Methodology

Methodologically, I regard the major contribution of this thesis to be the biologically relevant interpretation of including tumour fraction as a covariate in regression models during differential testing. All statistical frameworks already exist, and have even been used in similar ways before in [322]. However, the cited example was only limited to the simplest case (tumour fraction as the only covariate) and without in-depth interpretation of the regression coefficients. The reason why I think this is the major contribution is that the interpretation is not limited to DNA methylation data, but can be applied to all bulk quantitative assays including but not limited to RNA-seq, ChIP-seq, and chromatin accessibility assays.

However, while regression with tumour fraction as covariate allows studying both

tumour and non-tumour compartment, in practice it depends on an accurate tumour fraction estimation. According to my own experience in computational tumour content estimation, it appears that a lot of the CNV based or even SNV based purity estimates tend to overestimate tumour purity when the ground truth purity is low. The direct consequence for the regression is that while the modelled tumour methylation is a reasonably accurate, the estimated non-tumour methylation is often slightly higher than it should be especially when the tumour is hypomethylated in that region. Therefore, rather than ultra-sensitive methods for detecting tumour signals, I think in terms of molecular characterization using bulk assays, the field is in need of an unbiased tumour fraction estimation instead.

Another methodological innovation is the application of binomial mixture modelling in cfDNA methylation data over broad regions. Previous researchers have applied similar concept to utilize the locally phased methylation on each read, but limited their application to CpG-dense regions. In this thesis, I demonstrated that in the context of global hypomethylation in cancer, such methods can also be applied to broad regions, and show exciting preliminary results of potential clinical applications, such as early cancer detection and treatment monitoring. In this aspect, the use of binomial mixture modelling is completely orthogonal to any published methods for analysing cfDNA methylation data, and has the potential to be complementary to other cfDNA markers such as mutations, nucleosome positioning, and copy number alterations.

Lastly I have presented a copy number and purity aware method for classifying methylation subtypes in tumours. This was actually the part that I spent most of my effort on, but there are a few concerns that remain. In essence, the act of methylation subtyping is a type of unsupervised learning. Our cohort of $n = 23$ is rather insufficient for the discovery of convincing clusters, although I was still able to reproduce the clusters using publicly available data in a loose sense. In addition, the clustering is by nature under an important assumption that there really are molecular subtypes or cellular states in OAC that have intrinsically different DNA methylation, whereas in reality it could be the case that methylation is entirely

random, and no true methylation subtypes exist. The AIC metric was implemented to partially address this problem, but it cannot rule out the case of there being no subtypes. Wet lab validation would be required to solidify the findings.

6.2 Biology of aberrant cancer methylome

Thanks to realization of what it exactly means when performing regression using tumour fraction as covariate, I am able to model the average tumour methylation and non-tumour methylation, and subsequently visualize them using genome browser. This allowed me to have an intuitive exploration around loci of interest (see Fig. C.9 and Section 5.2.3 for example) to explore the relationship between DNA methylation and transcription in a gene-specific context.

In the introduction I have raised that we are beginning to understand that global hypomethylation in cancer is due to the failure in maintaining DNA methylation especially in late replicating regions, but we still don't have an explanation for focal hypermethylation. In this thesis, I have identified the positive correlation between late replication and focal hypermethylation, especially in genomic regions with CGI, H3K4me1, and H3K27me3 in normal cells of GI lineage. All three genomic elements have been mechanistically shown to antagonize DNA methylation marks. Together with the association with late replication, I therefore generated a testable hypothesis, where focal hypermethylation is a reactive change to the loss of protective marks, which in turn are secondary to failure of maintenance during late replication. Since histone marks can be lineage-specific, this provides an explanation for why certain CpG islands are preferentially methylated in some cancers but not others.

Another interesting aspect is the effect of DNA hypermethylation in histone-repressed genes. Although transcription is minimally affected, it may reflect a loss of epigenetic plasticity, from bivalent histone marks to repressive DNA methylation which are difficult to remove. Adding to this notion, it should be noted that the implicated genes are often developmental genes. Epigenetic plasticity is a rather abstract concept, but could perhaps be tested by the success rate of

reprogramming the cancer cells into pluripotent stem cells, with and without azacitidine or decitabine.

I look forward for these hypotheses to be tested, which could be performed using cancer cell lines directly or senescent cells from long term culture, similar to what Cruickshanks et al. [208] have done.

6.3 OAC methylation subtype

After defining the OAC methylation subtypes, I conducted extensive search for possible molecular mechanism that could explain the difference in DNA methylation. The single most striking feature is chromosomal instability and genome doubling in the hypomethylated tumour subtype (hypoT). Surprisingly, the relatively less hypomethylated tumour subtype (hyperT) demonstrates more aggressive features including an enrichment of MYC, E2F, and epithelial to mesenchymal transition (EMT) pathways. This is unexpected, because according to literature, hypomethylation should be proportional to proliferation. At present, there seems to be no straightforward explanation for the differences between the two subtypes.

The clinical implication of the methylation subtypes is uncertain. Based on our limited sample number and comparison with the TCGA cohort, patients with a dominant hypoT subtype in late stage disease might respond better to ICI+CTX than hyperT subtype. This information alone does not add much to clinical practice, because it appears that in late stage disease, the molecular tumour burden from cfDNA is a better prognostic indicator. More information regarding tumour response of subtypes in early stage cancer would be needed to judge the value of profiling the methylation subtype in OAC.

It is also a great limitation that, because only bulk sequencing data is available, I cannot tell for sure whether the methylation subtypes can coexist in the same tumour. To further investigate this matter, validation experiments using single-cell or mini-bulk methods would be needed. Additionally, if tumour cell lines or organoids can be derived from OAC patients, then the stability of the methylation subtypes could be measured, and functional assays could be performed.

6.4 Clinical relevance

The most apparent clinical relevance in this study is the analysis of cfDNA TAPS data. However, by the nature of the trial design, this study is largely exploratory, and any conclusions drawn from here should be considered preliminary. The immediate criticism of a near-perfect prediction of CB versus NCB patients from the cfDNA changes is that, the tumour tissue TAPS data were used to define tumour-specific loci, and subsequently applied to cfDNA in the same group of individuals. No external validation was available. Thus, one could argue that the encouraging results may be due to a tumour-informed analysis rather than a tumour-agnostic one, hence undermining the clinical utility. Moreover, although I have taken as many steps to avoid arbitrary cutoffs as possible, some cutoffs were still used, and I cannot deny the act of parameter tuning during the method development stage that ultimately led to the final results. Regardless, the discussions above should only affect the sensitivity and specificity of tumour detection, which can be optimized with more data and validation.

What is interesting and clear though is that in the patients who had NCB, tumour signal from cfDNA methylation increased after 4 weeks of ICI. One possibility is that the tumour cells are actively being killed, thus releasing tumour DNA into circulation. Patients with CB may have a more rapid response, and circulating tumour DNA (ctDNA) is already down-trending at 4 weeks of treatment, whereas those with NCB are still actively releasing ctDNA (Fig. 6.1).

Alternatively, the increased tumour signal in NCB could reflect a rapid increase of tumour proliferation after ICI, which may mean ICI is doing harm to patients. In view of this, I suggest that it would be an important clinical topic to figure out the dynamics of ctDNA following ICI-only treatment. More fine-grained longitudinal timepoints throughout the ICI-only treatment is required to distinguish between the two possibilities (delayed or slow tumour death versus increased tumour burden in NCB patients).

Another potentially clinically relevant finding is the association between TMB and tumour purity. I showed in Fig. 5.20 that except for the patient with MSI,

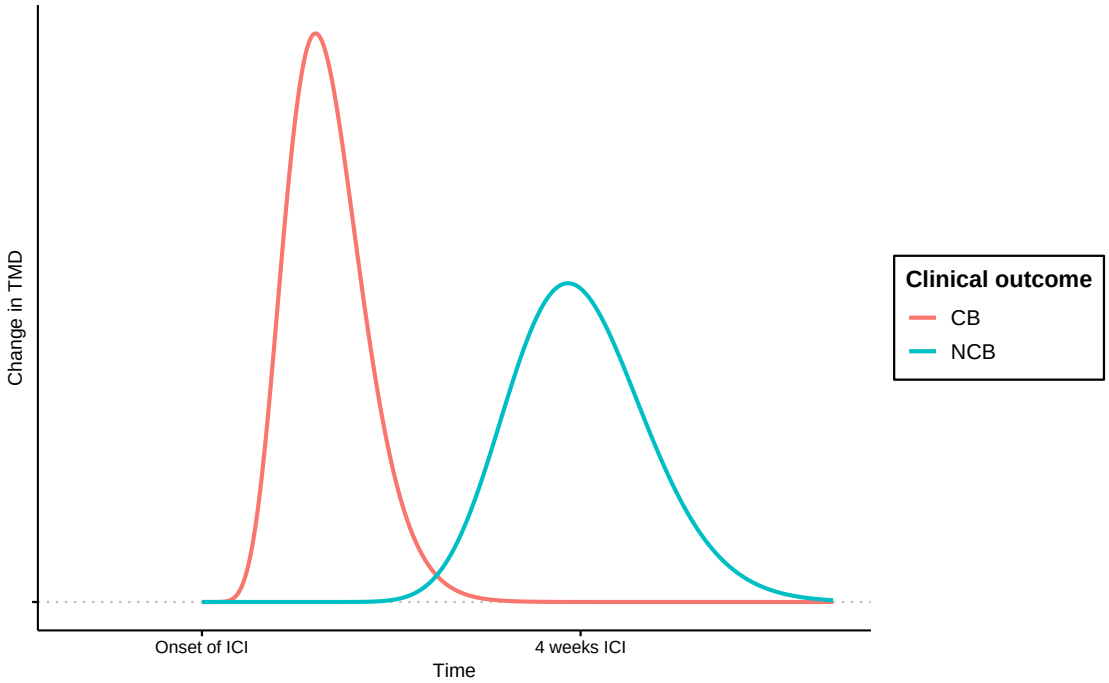


Fig. 6.1: Schematic diagram of changes in TMD after ICI in CB versus NCB patients.

there is an apparent linear relationship between TMB obtained using WGS and tumour purity. Since TMB is dependent on tumour purity and not the other way round, this suggests that the association between TMB and clinical outcome may actually be mediated through tumour purity of the biopsy. While this may seem absurd at first, tumour purity of biopsy could be reflecting the microenvironment of the tumour or the infiltrating ability of cancer cells. It has been previously described that low histologic purity in signet ring cell gastric cancer, which can have a poorly cohesive growth, is associated with metastasis and poor prognosis [323]. If this observation is true and applicable to pan-cancer types, then histologic purity or shallow WGS for CNA calling may be used to predict ICI-response instead of expensive high-depth sequencing.

6.5 Concluding remarks

I hope this thesis has contributed to the fields of cancer epigenetics and clinical oncology by providing new ways to use old tools.

Unfortunately, DNA methylation is not the only layer of epigenetic regulation, and as described throughout this work, coordinates with histone modifications in a complex manner. Currently, our understanding of the function of DNA methylation seems to have more exceptions than rules, especially in the cancer context. My attempts to approaching this subject from a top-down approach felt like walking on thin ice. However, such is the nature and beauty of academic research, as we are guided by curiosity and intuition, and we strive to design our experiments such that knowledge is gained regardless of failure or success. And if we failed to design such experiments, well, the failure itself has taught us something.

As much as I hate to admit, I have to say that my analyses on the DNA methylation landscape of OAC consisted of a lot of speculations and not many conclusions, especially on the functional role of DNA methylation in OAC. In the absence of reproducible experimental models, future studies should try to include as many multi-omic information as possible, including RNA-seq and either histone modifications or chromatin accessibility for a more tangible analysis of DNA methylation. It is also a shame that we are far from developing epigenetic drugs that can target specific genomic loci. Even if an important DMR is found, there is no way to specifically target it in clinical settings.

If significant progress were to be made in the biological aspects of cancer epigenetics, I suppose it would have to be about the regulatory mechanisms that led to different methylation phenotypes, rather than individual DMRs. The discovery of methylation subtypes based on global methylation pattern in this project could act as a foundation of such work. Meanwhile, DNA methylation shall continue to shine in the biomarker and diagnostics field, and hopefully increase the median survival in oesophageal cancers by effective screening and early treatment.

Appendices

A

R Markdown examples

A.1 Differential methylated region testing

See next page for embedded R Markdown document.

Examples for differentially methylated region testing

Phil Xie

Load packages and define helper functions

```
library(data.table)
library(ggplot2)
library(magrittr, include.only = '%>%')
library(ggtext)
library(ggthemes)
library(DSS)
source('ch-3/custom_DSS_func.R')
theme_set(theme_classic())

estBetaParams <- function(mu, var) {
  alpha <- ((1 - mu) / var - 1 / mu) * mu ^ 2
  beta <- alpha * (1 / mu - 1)
  return(list(alpha = alpha, beta = beta))
}

plot_arcsin_transformed <- function(x, coef) {
    c <- coef[1]
    m <- coef[2]
    y = m * x + c
    sapply(y, reverse_link)
}
```

Simulate data for example 1. Let the ground truth tumour methylation and non-tumour methylation be 0.1 and 0.8 respectively.

```
set.seed(123)
nsample <- 15
coverage <- round(rnorm(nsample, mean = 30, sd = 5))
tumour_fraction <- runif(nsample, min = 0.1, max = 1)

# model the tumour and non-tumour methylation using a beta distribution
tumour_avg_methylation <- 0.1
normal_avg_methylation <- 0.8
var <- 0.04
params <- estBetaParams(tumour_avg_methylation, var)
tumour_methylation <- rbeta(nsample, params$alpha, params$beta)
params <- estBetaParams(normal_avg_methylation, var)
non_tumour_methylation <- rbeta(nsample, params$alpha, params$beta)

# calculate the simulated methylation
simulated_methylation <-
    tumour_methylation * tumour_fraction +
    non_tumour_methylation * (1 - tumour_fraction)
```

```
# finally, simulate the sequencing data using a binomial distribution
simulated_mod <- rbinom(nsample, coverage, simulated_methylation)

# assemble everything into a data table
dt1 <- data.table(
    mod = simulated_mod,
    cov = coverage,
    tumour_fraction
)
```

DMR testing of tumour vs non-tumour. The design formula is:

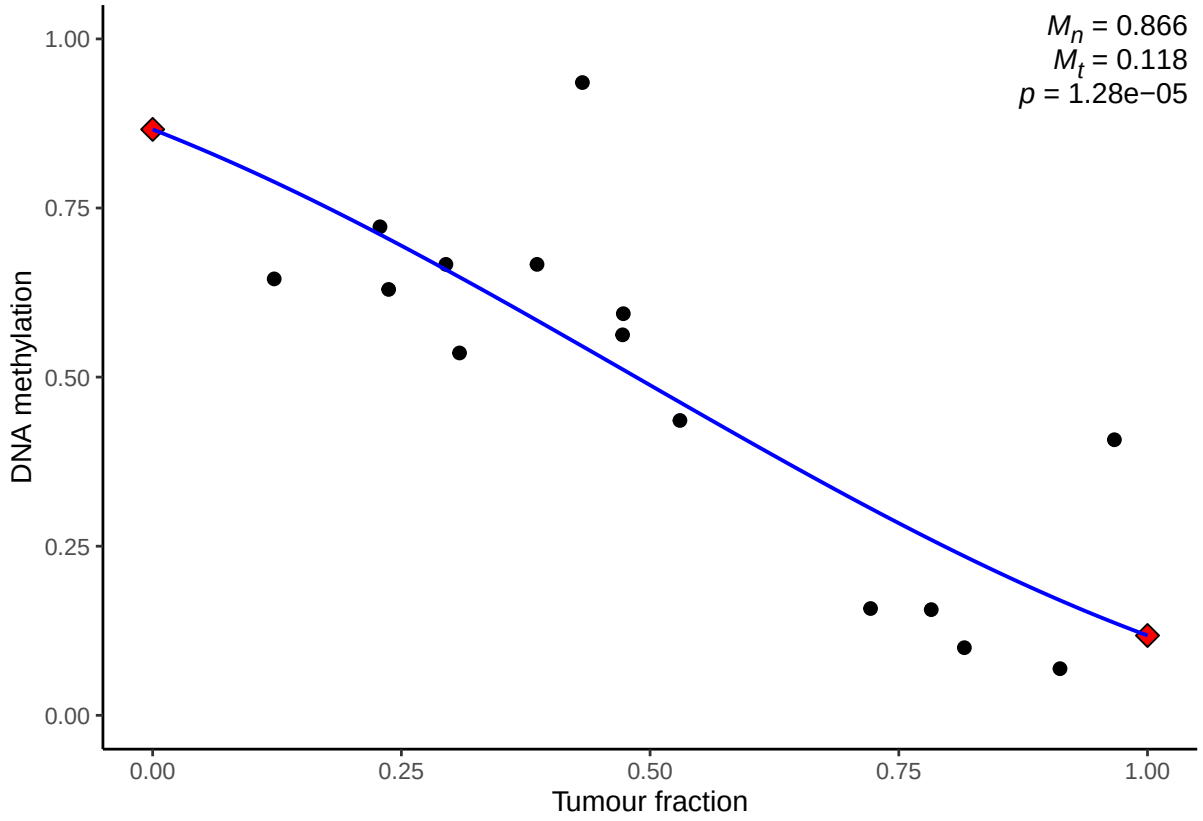
$M_o = \beta_0 + \beta_1 \lambda_t$

```
formula <- ~ tumour_fraction
terms <- list(normal = c(1, 0), tumour = c(1, 1))

invisible(capture.output(
    DMLfit <- DMLfit.multiFactor.light(dt1, design = dt1, formula = formula)
))
res <- DMLtest.multiFactor.contrast(
    DMLfit,
    Contrast = terms$tumour - terms$normal
)
res <- as.data.table(res)
res[, (names(terms)) :=
        lapply(terms, function(v) reverse_link(sum(DMLfit$fit$beta * v)))]

print(res)
```

```
##         stat         pval  maxcooks dispersion    loglik    normal    tumour
## 1: -4.364048 1.276774e-05 0.6364678  0.1200939 -6.039081 0.8662146 0.1180106
```

Visualize methylation and fitted model. Methylation can be plotted as a function of tumour fraction, as well as the fitted model:

```
dt1 %>%
    ggplot(aes(x = tumour_fraction, y = mod / cov)) +
    geom_point(size = 2) +
    annotate('point', x = c(0, 1), y = c(res$normal, res$tumour),
             size = 3, shape = 23, fill = 'red') +
    annotate('richtext', x = Inf, y = Inf, hjust = 1, vjust = 1,
             fill = NA, label.color = NA,
             label = glue::glue('
             *M<sub>n</sub>* = {signif(res$normal, 3)}<br>
             *M<sub>t</sub>* = {signif(res$tumour, 3)}<br>
             *p* = {signif(res$pval, 3)}
             ')) +
    geom_function(
        col = 'blue', linewidth = 0.7,
        fun = ~ plot_arcsin_transformed(.x, DMLfit$fit$beta)
    ) +
    expand_limits(x = c(0, 1), y = c(0, 1)) +
    xlab('Tumour fraction') + ylab('DNA methylation')
```



Interpretation. The estimated tumour methylation M_t is 0.12, and non-tumour methylation M_n is 0.87. Both are reasonable estimates of the ground truth of 0.1 and 0.8 respectively. The p value is 1.28e-5, meaning there is a significant difference between the tumour and non-tumour methylation.

Simulate data for example 2. Let there be tumour subtypes A and B, where subtype A is exactly the same as in example 1. Let the ground truths for tumour and non-tumour methylation in subtype B be 0.8 and 0.6 respectively.

```
set.seed(234)
nsample <- 15
coverage <- round(rnorm(nsample, mean = 30, sd = 5))
tumour_fraction <- runif(nsample, min = 0.1, max = 1)

# model the tumour and non-tumour methylation using a beta distribution
tumour_avg_methylation <- 0.8
normal_avg_methylation <- 0.6
var <- 0.04
params <- estBetaParams(tumour_avg_methylation, var)
tumour_methylation <- rbeta(nsample, params$alpha, params$beta)
params <- estBetaParams(normal_avg_methylation, var)
non_tumour_methylation <- rbeta(nsample, params$alpha, params$beta)

# calculate the simulated methylation
simulated_methylation <-
    tumour_methylation * tumour_fraction +
    non_tumour_methylation * (1 - tumour_fraction)
```

```
# finally, simulate the sequencing data using a binomial distribution
simulated_mod <- rbinom(nsample, coverage, simulated_methylation)

# assemble everything into a data table
dt2 <- data.table(
    mod = simulated_mod,
    cov = coverage,
    tumour_fraction
)

# combine with subtype A
dt <- rbindlist(list(A = dt1, B = dt2), idcol = 'subtype')
```

DMR testing of tumour subtypes. Here we can include subtype as an additional covariate in the model design. Let $S = 0$ for subtype A, and $S = 1$ for subtype B. Since DNA methylation of both the tumour and the microenvironment can be different between the two subtypes, we need to include interaction terms for both the slope and intercept. Hence, the design formula is:

$M_o = \beta_0 + \beta_1\lambda_t + \beta_2 S + \beta_3\lambda_t S$

There are many possible comparisons that can be performed in this case, including:

(1) Tumour subtype A versus tumour subtype B
(2) Non-tumour subtype A versus non-tumour subtype B
(3) Tumour subtype A versus non-tumour subtype A
(4) Tumour subtype B versus non-tumour subtype B

Here, I will perform (1) and (2) as examples:

```
formula <- ~ tumour_fraction * subtype
terms <- list(
    normal_A = c(1, 0, 0, 0),
    tumour_A = c(1, 1, 0, 0),
    normal_B = c(1, 0, 1, 0),
    tumour_B = c(1, 1, 1, 1)
)

invisible(capture.output(
    DMLfit <- DMLfit.multiFactor.light(dt, design = dt, formula = formula)
))

tmp <- terms[c('tumour_A', 'tumour_B')]
res1 <- DMLtest.multiFactor.contrast(
    DMLfit,
    Contrast = tmp[[1]] - tmp[[2]]
)
res1 <- as.data.table(res1)
res1[, (names(tmp)) :=
        lapply(tmp, function(v) reverse_link(sum(DMLfit$fit$beta * v)))]

tmp <- terms[c('normal_A', 'normal_B')]
res2 <- DMLtest.multiFactor.contrast(
    DMLfit,
    Contrast = tmp[[1]] - tmp[[2]]
)
res2 <- as.data.table(res2)
```

```
res2[, (names(tmp)) :=
        lapply(tmp, function(v) reverse_link(sum(DMLfit$fit$beta * v)))]

# DMR result: tumour A vs tumour B
print(res1)
```

```
##         stat         pval  maxcooks dispersion    loglik  tumour_A  tumour_B
## 1: -3.718374 0.0002005094 0.2280641  0.2042948 -17.40759 0.1188835 0.7689633
```

```
# DMR result: non-tumour A vs non-tumour B
print(res2)
```

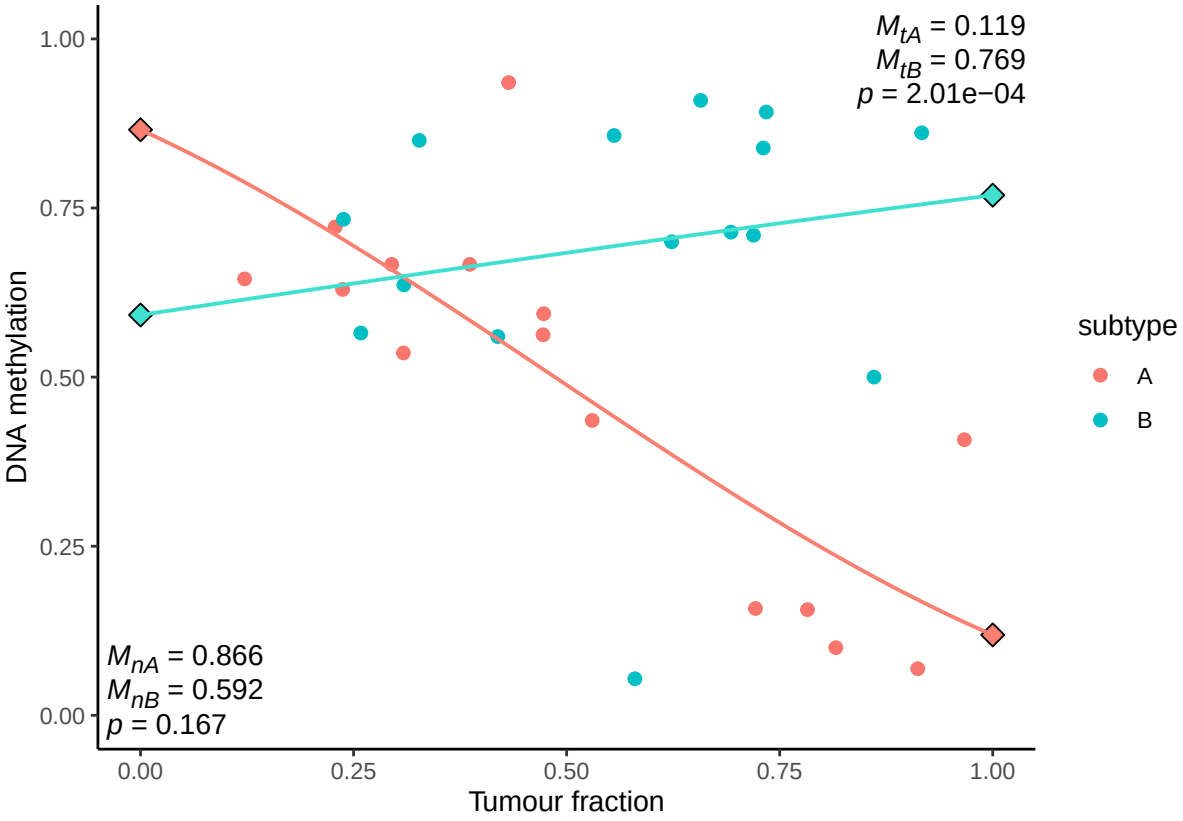
```
##        stat      pval  maxcooks dispersion    loglik  normal_A  normal_B
## 1: 1.381303 0.1671859 0.2280641  0.2042948 -17.40759 0.8656684 0.5918532
```

Visualize methylation and fitted model. From the design formula, we have

$M_o = \beta_0 + \beta_1 \lambda_t$ for subtype A, where $S = 0$ $M_o = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)\lambda_t$ for subtype B, where $S = 1$

Therefore by treating S as a fixed variable, we may plot methylation as a function of tumour fraction for subtype A and subtype B respectively.

```
dt %>%
    ggplot(aes(x = tumour_fraction, y = mod / cov, color = subtype)) +
    geom_point(size = 2) +
    annotate('point', x = c(0, 1), y = c(res2$normal_A, res1$tumour_A),
             size = 3, shape = 23, fill = 'salmon') +
    annotate('point', x = c(0, 1), y = c(res2$normal_B, res1$tumour_B),
             size = 3, shape = 23, fill = 'turquoise') +
    geom_function(
        linewidth = 0.7, col = 'salmon',
        fun = ~ plot_arcsin_transformed(.x, DMLfit$fit$beta)
    ) +
    geom_function(
        linewidth = 0.7, col = 'turquoise',
        fun = ~ plot_arcsin_transformed(.x, rowSums(matrix(DMLfit$fit$beta, nrow = 2)))
    ) +
    annotate('richtext', x = Inf, y = Inf, hjust = 1, vjust = 1,
             fill = NA, label.color = NA,
             label = glue::glue('
             *M<sub>tA</sub>* = {signif(res1$tumour_A, 3)}<br>
             *M<sub>tB</sub>* = {signif(res1$tumour_B, 3)}<br>
             *p* = {format(signif(res1$pval, 3), scientific = T)}
             ')) +
    annotate('richtext', x = -Inf, y = -Inf, hjust = 0, vjust = 0,
             fill = NA, label.color = NA,
             label = glue::glue('
             *M<sub>nA</sub>* = {signif(res2$normal_A, 3)}<br>
             *M<sub>nB</sub>* = {signif(res2$normal_B, 3)}<br>
             *p* = {signif(res2$pval, 3)}
             ')) +
    expand_limits(x = c(0, 1), y = c(0, 1)) +
    xlab('Tumour fraction') + ylab('DNA methylation')
```



Interpretation. The estimated tumour and non-tumour methylation values for subtype A are 0.12 and 0.87 respectively. Note that the estimated values are slightly different from those in example 1, because sample subtypes A and B are modeled together, resulting in a slightly different dispersion estimate. The corresponding values are 0.77 and 0.59 for subtype B, which are good estimates of the ground truth values of 0.8 and 0.6.

The p value for subtype A tumour versus subtype B tumour comparison is 2e-4, suggesting a significant difference. The p value for subtype A non-tumour versus subtype B non-tumour is 0.16, suggesting that there is not enough power to reject the null hypothesis at the given sequencing coverage, sample size, and methylation difference.

A.2 Binomial mixture modelling of read-based methylation

See next page for embedded R Markdown document.

Binomial mixture modeling

Phil Xie

Load packages and define helper functions

```
library(data.table)
library(ggplot2)
library(magrittr, include.only = '%>%')
library(ggthemes)
library(flexmix)
theme_set(theme_classic())

# defining the function to perform mixture modeling
fitBinoMix <- function(input, k, nrep = 3, minprior = 0.001) {

  design <- cbind(mod, unmod) ~ 1 # hardcoded variable names
  model <- FLXMRglm(family = "binomial")
  # cem.em is the fastest; minprior defines the limit of detection
  # default control params can be viewed by running new('FLXcontrol')
  m1 <- initFlexmix(
    design, data = input, k = k, model = model, nrep = nrep,
    verbose = F, unique = T, init = list(name = 'cem.em'),
    control = list(minprior = minprior, iter.max = 100 / minprior)
  )
  if (length(k) > 1) {
    converged <- sapply(m1@models, function(x) x@converged)
    sel <- names(which.min(AIC(m1)[converged]))
    bestm <- `if`(is.null(sel), NA, getModel(m1, which = sel))
  } else {
    bestm <- m1
    if (!bestm@converged) bestm <- NA
  }
  return(bestm)
}

# clusters are named by descending order of cluster size
getParams <- function(model) {
  if (suppressWarnings(is.na(model))) return(list(
      cluster = as.integer(NA), size = as.integer(NA), mu = as.numeric(NA)
  ))
  reorder <- order(-model@size)
  mu <- plogis(parameters(model))

  list(
    cluster = seq_along(reorder),
    size = model@size[reorder],
    mu = mu[reorder]
```

```
  )
}

# generating random data based on a few key parameters
simulateData <- function(n = 1e4, nCG = 3, tf = 0.1,
                         tbeta = 0.1, nbeta = 0.7, seed = 123) {
  # n : number of reads
  # nCG : number of CpG on each read, modelled using negative binomial
  # tf : ground truth tumour fraction
  # tbeta : ground truth tumour methylation
  # nbeta : ground truth non-tumour methylation

  set.seed(seed)
  reads <- rnbinom(n, nCG, mu = nCG) # number of CpG on each read
  treads <- sample(reads, n*tf)
  nreads <- sample(reads, n*(1-tf))

  tmreads <- data.table(
      cov = treads,
      mod = as.integer(sapply(treads, function(x) rbinom(1, x, tbeta)))
  )
  nmreads <- data.table(
      cov = nreads,
      mod = as.integer(sapply(nreads, function(x) rbinom(1, x, nbeta)))
  )
  dt <- rbindlist(list(tumo = tmreads, norm = nmreads), idcol = 'ident')
  dt[, unmod := cov - mod]
  dt <- dt[cov != 0]
  return(dt)
}
```

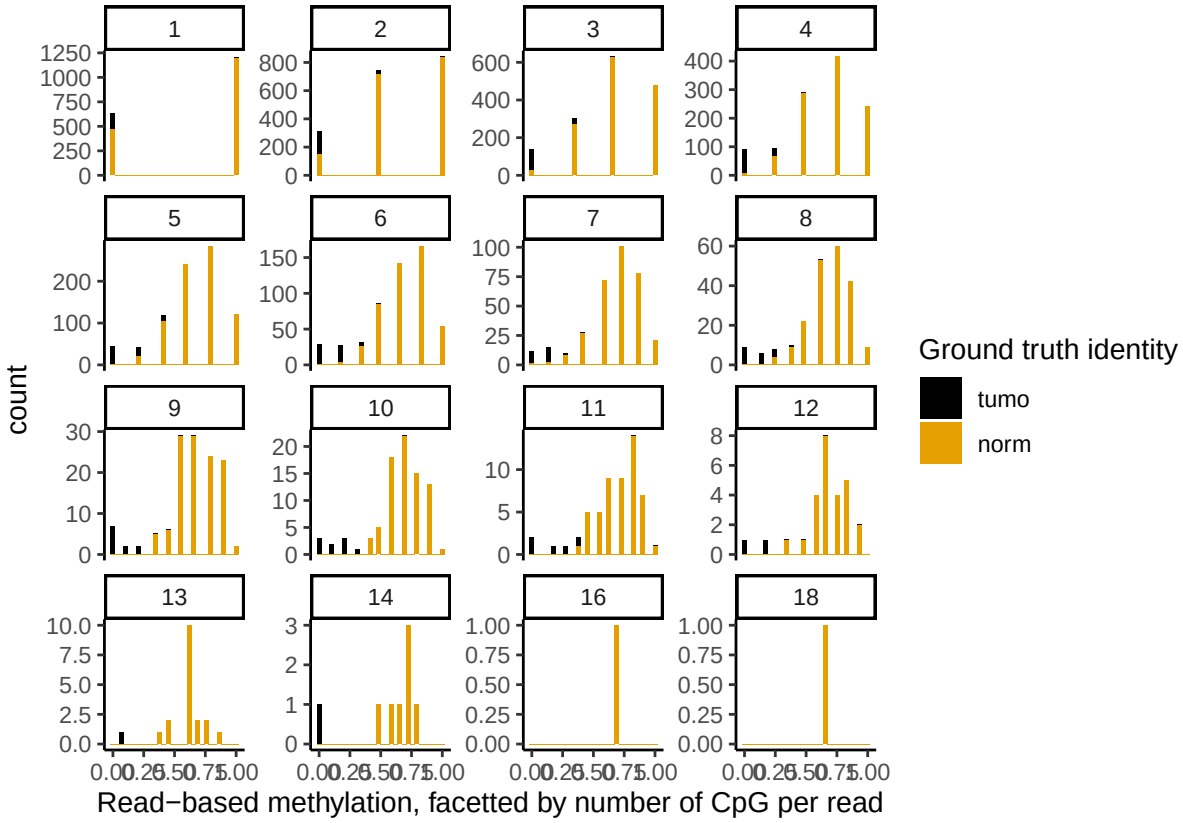
Model performance using simulated data

```
minprior <- 0.001 # cutoff for lower limit of detection

# Simulate some data
dt <- simulateData(tf = 0.1, tbeta = 0.1, nbeta = 0.7)

set.seed(123)
bestm <- fitBinoMix(dt, k = 1:3, minprior = minprior)
getParams(bestm)
```

```
## $cluster
## [1] 1 2
##
## $size
##    2    1
## 8254  447
##
## $mu
## Comp.2.coef.(Intercept) Comp.1.coef.(Intercept)
##               0.6982184               0.0969191
```

Visualize the distribution of read-based methylation.

```
dt %>% ggplot(aes(x = mod/cov, fill = forcats::fct_rev(ident))) +
    geom_histogram(bins = 30) +
    scale_fill_colorblind(name = 'Ground truth identity') +
    facet_wrap(~ cov, scales = 'free_y') +
    xlab('Read-based methylation, facetted by number of CpG per read')
```



The model performs well for tumour fraction of 10%. Can we push it down to 1%?

```
# Simulate some data
dt <- simulateData(tf = 0.01, tbeta = 0.1, nbeta = 0.7)

set.seed(123)
bestm <- fitBinoMix(dt, k = 1:3, minprior = minprior)
getParams(bestm)
```

```
## $cluster
## [1] 1 2
##
## $size
##    1    2
## 8642   40
##
## $mu
## Comp.1.coef.(Intercept) Comp.2.coef.(Intercept)
##              0.70228800              0.07882454
```

Can we push it down to 0.1%?

```r
minprior <- 0.001 # cutoff for lower limit of detection
# Simulate some data
dt <- simulateData(tf = 0.001, tbeta = 0.1, nbeta = 0.7)

set.seed(123)
bestm <- fitBinoMix(dt, k = 1:3, minprior = minprior)
getParams(bestm)
```

```
## $cluster
## [1] 1
##
## $size
##    1
## 8684
##
## $mu
## Comp.1.coef.(Intercept)
##               0.6977431
```

We can't. But if we increase the number of reads, we can recover it:

```r
minprior <- 0.001 # cutoff for lower limit of detection
# Simulate some data
dt <- simulateData(tf = 0.001, tbeta = 0.1, nbeta = 0.7, n = 1e5)

set.seed(123)
bestm <- fitBinoMix(dt, k = 1:3, minprior = minprior)
getParams(bestm)
```

```
## $cluster
## [1] 1 2
##
## $size
##     2     1
## 87281     8
##
## $mu
## Comp.2.coef.(Intercept) Comp.1.coef.(Intercept)
##               0.7010733               0.4027919
```

Although the estimated mean would be far from the ground truth (0.4 versus 0.1). It has also been noted that the cluster size of the minor cluster is consistently underestimated.

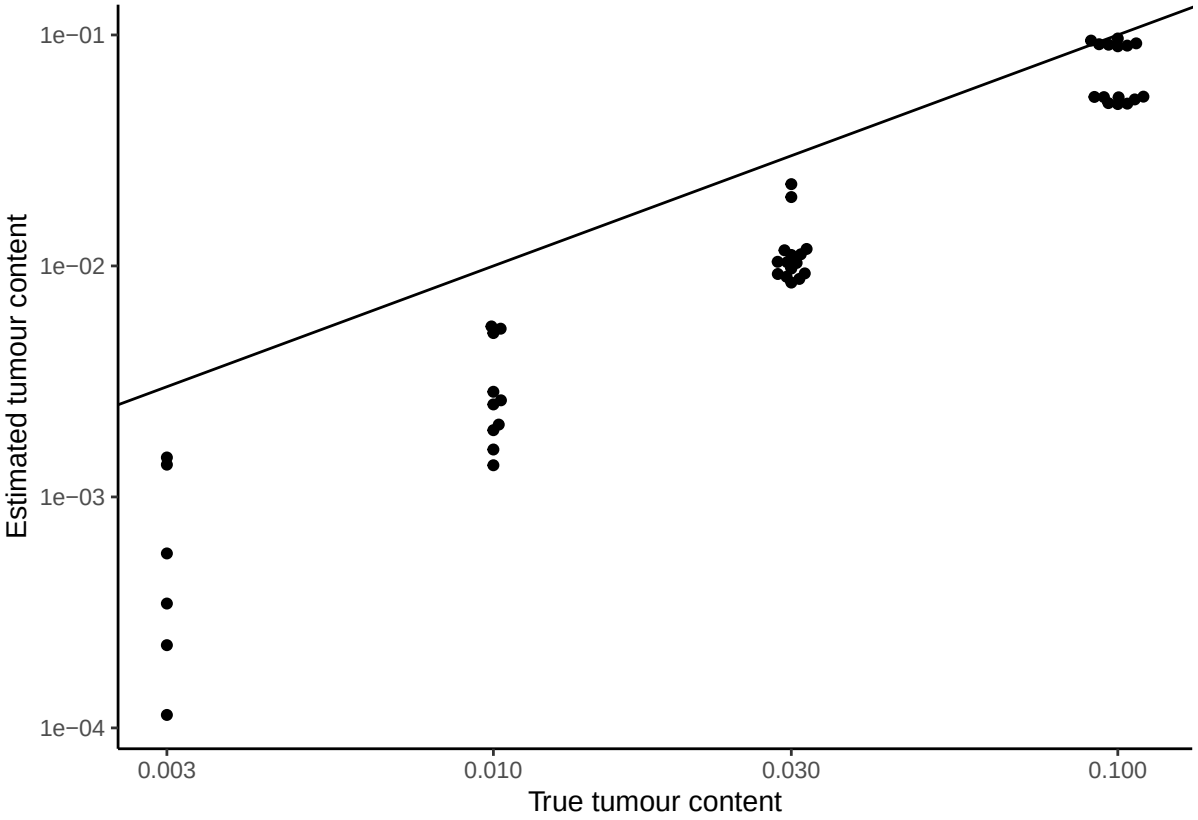
Finally, create a more systematic comparison

```r
minprior <- 0.001 # cutoff for lower limit of detection
nreads <- 1e4
nrep <- 15
# Simulate some data
mixtureprob <- rep(c(0, 0.001, 0.003, 0.01, 0.03, 0.1), nrep)
dt <- lapply(seq_along(mixtureprob), function(i) {
    seed <- i
    prob <- mixtureprob[i]
    x <- simulateData(tf = prob, tbeta = 0.1, nbeta = 0.7, n = nreads, seed = i)
    x$true_tf <- prob
    x$exmpt <- seed
    return(x)
```

```
})

outparam <- rbindlist(dt)[, getParams(fitBinoMix(.SD, k = 1:3)),
                          by = .(true_tf, exmpt)]
```

Visualization of the estimated versus ground truth tumour content. Straight line represents y = x. Note that for low tumour content, mixture modelling consistently underestimates the tumour content. This is probably an effect due to model selection using information criteria such as AIC, which penalizes fitting additional parameters.

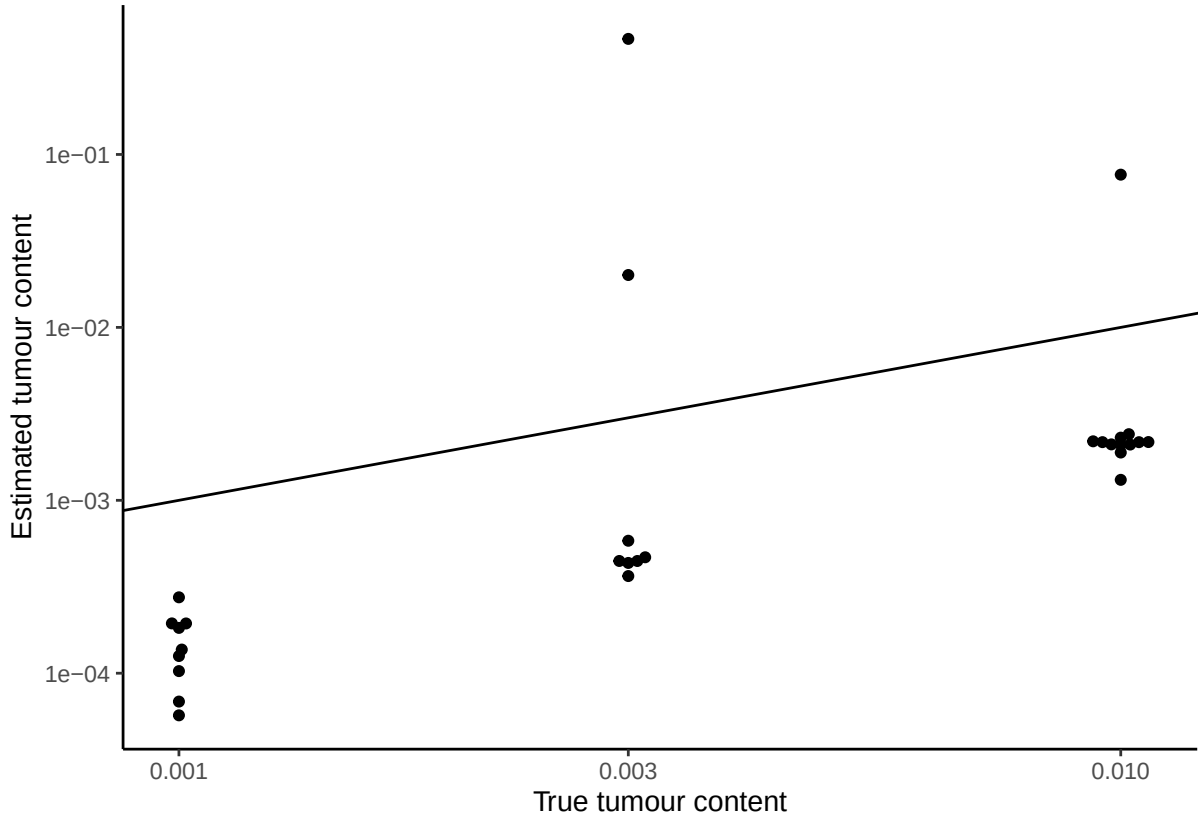
```
outparam <- outparam[size != 0]
outparam[, mix := size / sum(size), by = exmpt]
outparam[, nclust := .N, by = exmpt]

outparam %>%
    .[true_tf != 0] %>% # removed to avoid NaN on log x-axis
    .[nclust == 2, .(est_mix = min(mix)), by = .(true_tf, exmpt)] %>%
    ggplot(aes(x = true_tf, y = est_mix, group = true_tf)) +
    ggbeeswarm::geom_beeswarm() +
    geom_abline(slope = 1, intercept = 0) +
    scale_x_log10() + scale_y_log10() +
    xlab('True tumour content') + ylab('Estimated tumour content')
```



Since all mixture modelling for 0.1% tumour content failed at 1e4 sequencing reads, the modelling is performed again at 1e5 sequencing reads for the low tumour content samples.

```r
minprior <- 0.001 # cutoff for lower limit of detection
nreads <- 1e5
# Simulate some data
mixtureprob <- rep(c(0, 0.001, 0.003, 0.01), nrep)
dt2 <- lapply(seq_along(mixtureprob), function(i) {
    seed <- i
    prob <- mixtureprob[i]
    x <- simulateData(tf = prob, tbeta = 0.1, nbeta = 0.7, n = nreads, seed = i)
    x$true_tf <- prob
    x$exmpt <- seed
    return(x)
})

outparam2 <- rbindlist(dt2)[, getParams(fitBinoMix(.SD, k = 1:3)),
                            by = .(true_tf, exmpt)]
```

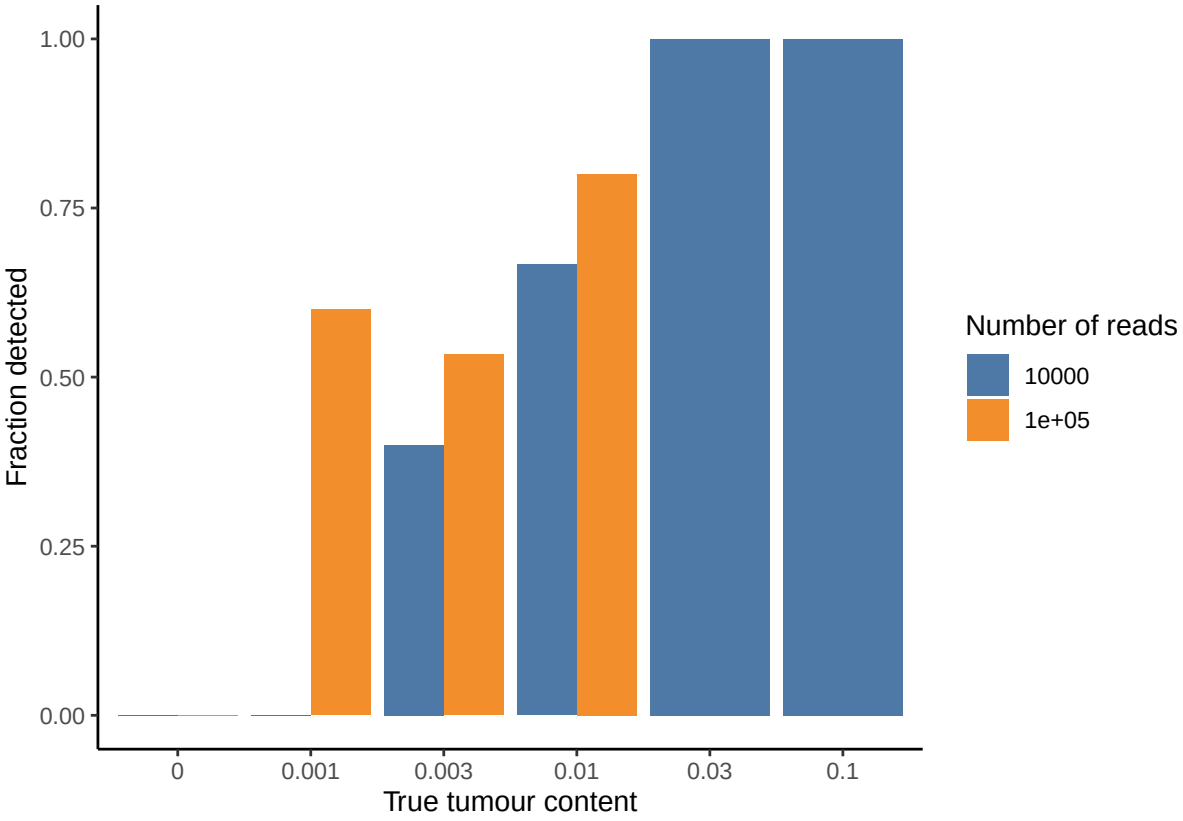
```r
outparam2 <- outparam2[size != 0]
outparam2[, mix := size / sum(size), by = exmpt]
outparam2[, nclust := .N, by = exmpt]

outparam2 %>%
    .[true_tf != 0] %>% # removed to avoid NaN on log x-axis
    .[nclust == 2, .(est_mix = min(mix)), by = .(true_tf, exmpt)] %>%
    ggplot(aes(x = true_tf, y = est_mix, group = true_tf)) +
    ggbeeswarm::geom_beeswarm() +
    geom_abline(slope = 1, intercept = 0) +
    scale_x_log10() + scale_y_log10() +
    xlab('True tumour content') + ylab('Estimated tumour content')
```



Below reports the fraction of simulations where a second component is detected, given a fixed number of reads. This suggests that on average, 80% of tumour specific loci, each covered by 1e5 reads, are able to detect tumour fraction at 0.1%. Meanwhile, the false positive rate remains 0%.

```
rbindlist(list('1e4' = outparam, '1e5' = outparam2), idcol = 'num_reads') %>%
    .[, .(frac_detected = sum(cluster == 2) / nrep),
      by = .(true_tf, num_reads = as.numeric(num_reads))] %T>%
    print %>%
    ggplot(aes(x = factor(true_tf), y = frac_detected, fill = factor(num_reads))) +
    geom_col(position = 'dodge') +
    scale_fill_tableau(name = 'Number of reads') +
    xlab('True tumour content') + ylab('Fraction detected')
```

```
##     true_tf num_reads frac_detected
##  1:   0.000     1e+04     0.0000000
##  2:   0.001     1e+04     0.0000000
##  3:   0.003     1e+04     0.4000000
##  4:   0.010     1e+04     0.6666667
##  5:   0.030     1e+04     1.0000000
##  6:   0.100     1e+04     1.0000000
##  7:   0.000     1e+05     0.0000000
##  8:   0.001     1e+05     0.6000000
##  9:   0.003     1e+05     0.5333333
## 10:   0.010     1e+05     0.8000000
```


1.00
0.75
0.50
0.25
0.00
Fraction detected
0
0.001
0.003
0.01
0.03
0.1
True tumour content
Number of reads
10000
1e+05

A.3 Tumour subtyping

See next page for embedded R Markdown document.

Obtaining tumour fraction and copy number aware methylation status

Phil Xie

Load packages and define helper functions

```
library(data.table)
library(ggplot2)
library(magrittr, include.only = '%>%')
library(ggtext)
library(ggthemes)
library(DSS)
source('ch-3/custom_DSS_func.R')
theme_set(theme_classic())

estBetaParams <- function(mu, var) {
  alpha <- ((1 - mu) / var - 1 / mu) * mu ^ 2
  beta <- alpha * (1 / mu - 1)
  return(list(alpha = alpha, beta = beta))
}

plot_arcsin_transformed <- function(x, coef) {
    c <- coef[1]
    m <- coef[2]
    y = m * x + c
    sapply(y, reverse_link)
}

plot_logit_transformed <- function(x, coef) {
    c <- coef[1]
    m <- coef[2]
    y = m * x + c
    plogis(y)
}
```

Simulate data from 2 tumour subtypes. Let the ground truth tumour non-tumour methylation in subtype A be 0.1 and 0.8 respectively.

```
set.seed(123)
nsample <- 15
coverage <- round(rnorm(nsample, mean = 30, sd = 5))
tumour_fraction <- runif(nsample, min = 0.1, max = 1)

# model the tumour and non-tumour methylation using a beta distribution
tumour_avg_methylation <- 0.1
normal_avg_methylation <- 0.8
var <- 0.04
```

```
params <- estBetaParams(tumour_avg_methylation, var)
tumour_methylation <- rbeta(nsample, params$alpha, params$beta)
params <- estBetaParams(normal_avg_methylation, var)
non_tumour_methylation <- rbeta(nsample, params$alpha, params$beta)

# calculate the simulated methylation
simulated_methylation <-
    tumour_methylation * tumour_fraction +
    non_tumour_methylation * (1 - tumour_fraction)

# finally, simulate the sequencing data using a binomial distribution
simulated_mod <- rbinom(nsample, coverage, simulated_methylation)

# assemble everything into a data table
dt1 <- data.table(
    mod = simulated_mod,
    cov = coverage,
    tumour_fraction
)
```

Let the ground truths for tumour and non-tumour methylation in subtype B be 0.8 and 0.6 respectively.

```
set.seed(234)
nsample <- 15
coverage <- round(rnorm(nsample, mean = 30, sd = 5))
tumour_fraction <- runif(nsample, min = 0.1, max = 1)

# model the tumour and non-tumour methylation using a beta distribution
tumour_avg_methylation <- 0.8
normal_avg_methylation <- 0.6
var <- 0.04
params <- estBetaParams(tumour_avg_methylation, var)
tumour_methylation <- rbeta(nsample, params$alpha, params$beta)
params <- estBetaParams(normal_avg_methylation, var)
non_tumour_methylation <- rbeta(nsample, params$alpha, params$beta)

# calculate the simulated methylation
simulated_methylation <-
    tumour_methylation * tumour_fraction +
    non_tumour_methylation * (1 - tumour_fraction)

# finally, simulate the sequencing data using a binomial distribution
simulated_mod <- rbinom(nsample, coverage, simulated_methylation)

# assemble everything into a data table
dt2 <- data.table(
    mod = simulated_mod,
    cov = coverage,
    tumour_fraction
)

# combine with subtype A
dt <- rbindlist(list(A = dt1, B = dt2), idcol = 'subtype')
```

Tumour subtyping

We will pretend that we do not know the ground truth subtypes. The gist of my subtyping approaches would be that, instead of using a overdispersed model, we use multiple non-overdispersed model and see what fits best.

I have two approaches; the first is more crude but faster, the second in theory handles certain situation better, but is much slower.

Approach 1: binomial test. First, build a simple model with only tumour fraction as covariate.

```
formula <- ~ tumour_fraction
terms <- list(normal = c(1, 0), tumour = c(1, 1))

invisible(capture.output(
    DMLfit_step1 <- DMLfit.multiFactor.light(dt, design = dt, formula = formula)
))
res_step1 <- DMLtest.multiFactor.contrast(
    DMLfit_step1,
    Contrast = terms$tumour - terms$normal
)
res_step1 <- as.data.table(res_step1)
res_step1[, (names(terms)) :=
        lapply(terms, function(v) reverse_link(sum(DMLfit_step1$fit$beta * v)))]

print(res_step1)
```

```
##         stat      pval  maxcooks dispersion    loglik    normal    tumour
## 1: -1.572392 0.1158597 0.1962206  0.3326341 -24.76307 0.7718446 0.4266289
```

Next, based on the regression model coefficients, get the predicted methylation for each tumour at the corresponding tumour purity.

Then, it is assumed that sample methylation would follow a binomial distribution with the expected mean being the predicted mean from above.

A statistical test can then be performed to see how likely it is to obtain the methylation sequencing counts given the predicted methylation. Using a p-value cutoff of 0.05, each data point can then be classified as significantly above, below, or not significantly different from the overall mean.

Note that in reality the overall methylation is a mixture of tumour and non-tumour methylation, which means that strictly speaking the overall methylation is not a single binomial distribution, but instead is a mixture of 2 binomial distributions. There is therefore a flaw in the mathematical rigor when performing the binomial test.

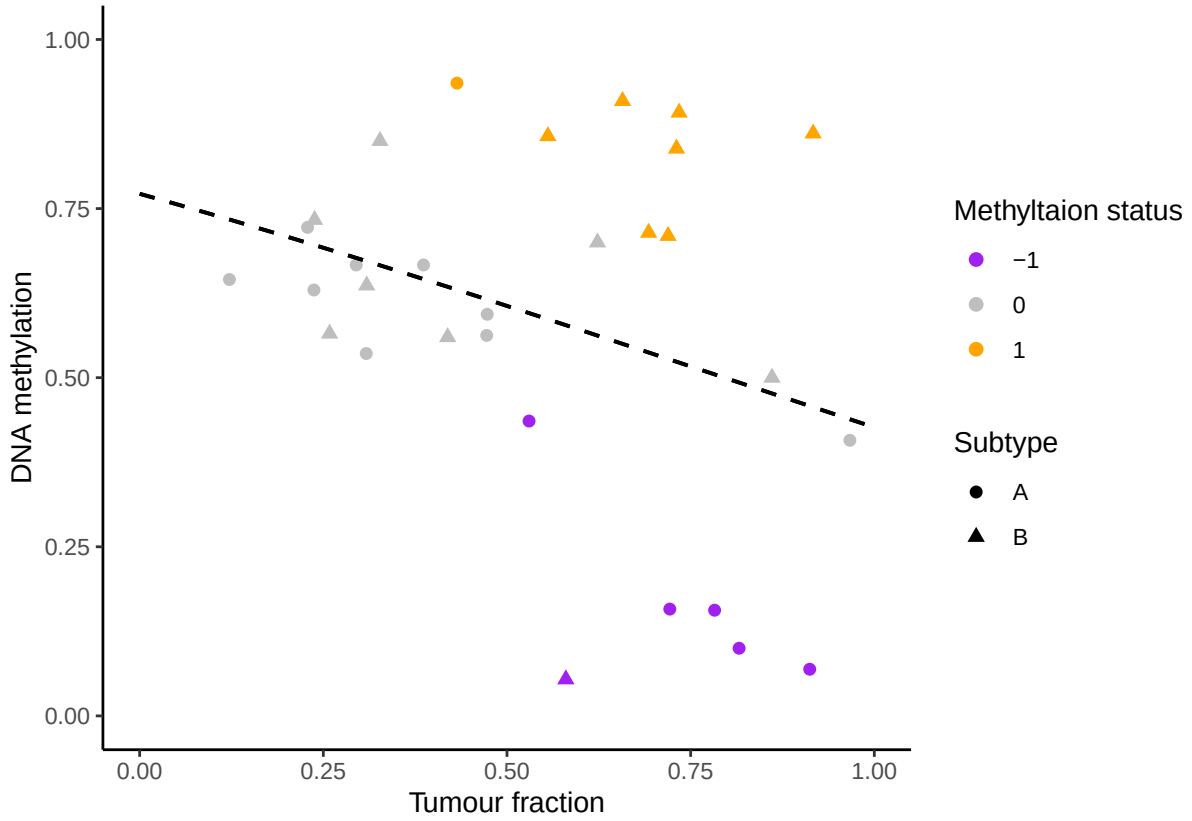
```
dt[, pred := sapply(DMLfit_step1$X %*% t(DMLfit_step1$fit$beta), reverse_link)]
# now, assume binomial instead of beta-binomial distribution
# calculate the p value for the observed methylation
# binom_pval is pre-defined as a numeric variable for coding reasons
dt[, binom_pval := as.numeric(NA)]
dt[, binom_pval := binom.test(mod, cov, pred)$p.value, by = 1:nrow(dt)]
# binarize the unlikely events into "higher" or "lower" than expected
dt[binom_pval < 0.05, binarize := sign((mod / cov) - pred)]
setnafill(dt, fill = 0, cols = 'binarize')
```

Plot the graph to see if the predictions are reasonable. Shape of the dot corresponds to the ground truth tumour subtype, whereas colour of the dot represents the subtype prediction. The black dotted line represents the overall mean, as modelled with only tumour fraction as covariate.

```
dt[binarize != 0, table(subtype, binarize)]
```

```
##         binarize
## subtype -1 1
##       A  5 1
##       B  1 7
```

```
overallCoef <- rowSums(matrix(DMLfit_step1$fit$beta, nrow = 2))
dt %>%
    ggplot(aes(x = tumour_fraction, y = mod / cov, shape = subtype,
               color = factor(binarize), group = subtype)) +
    geom_point(size = 2) +
    geom_function(
        linewidth = 0.7, col = 'black', linetype = 2,
        fun = ~ plot_arcsin_transformed(.x, overallCoef)
    ) +
    scale_color_manual(values = c('purple', 'grey', 'orange'),
                       name = 'Methyltaion status') +
    scale_shape(name = 'Subtype') +
    expand_limits(x = c(0, 1), y = c(0, 1)) +
    xlab('Tumour fraction') + ylab('DNA methylation')
```



Out of 30 samples, 14 samples were classified to be significantly different from the overall mean, and most samples were classified correctly in the simulated example. When this is repeated for every loci genome wide, we may then perform hierarchical clustering on the digital matrix in a fashion similar to those in previous publications.

However, in scenarios where tumour methylation of subtype A is higher than subtype B, AND non-tumour

methylation of subtype B is higher than subtype A, this approach may lead to inaccurate results. High purity subtype A samples would be classified as higher than overall mean, whereas low purity subtype A samples would be classified as lower than overall mean. This brings us to the second approach.

Approach 2: binomial regression mixture modelling An alternative is to build a mixture model using binomial regression of n components. Since we are mostly interested in the binary methylation status, we choose $n \leq 2$.

```
library(flexmix)
```

```
## Warning: package 'flexmix' was built under R version 4.2.3
```

```
## Loading required package: lattice
```

```
set.seed(234)
mdl <- initFlexmix(
    cbind(mod, I(cov - mod)) ~ tumour_fraction, data = dt, k = 1:2,
    model = FLXglm(family = 'binomial'), nrep = 3, verbose = F, unique = T,
    init = list(name = 'cem.em')
) %>% getModel('AIC')

# set confidence cutoff
# only data with good posterior probability will be classified
# also, order the clusters so the one with lower tumour methylation goes first
clus <- clusters(mdl)
confidence_cut <- 0.95
clus[rowSums(mdl@posterior$scaled > confidence_cut) == 0] <- NA
binarizedrank <- c(-1, 1)[rank(plogis(colSums(parameters(mdl))))]
clus <- binarizedrank[clus]

table(dt$subtype, clus)
```

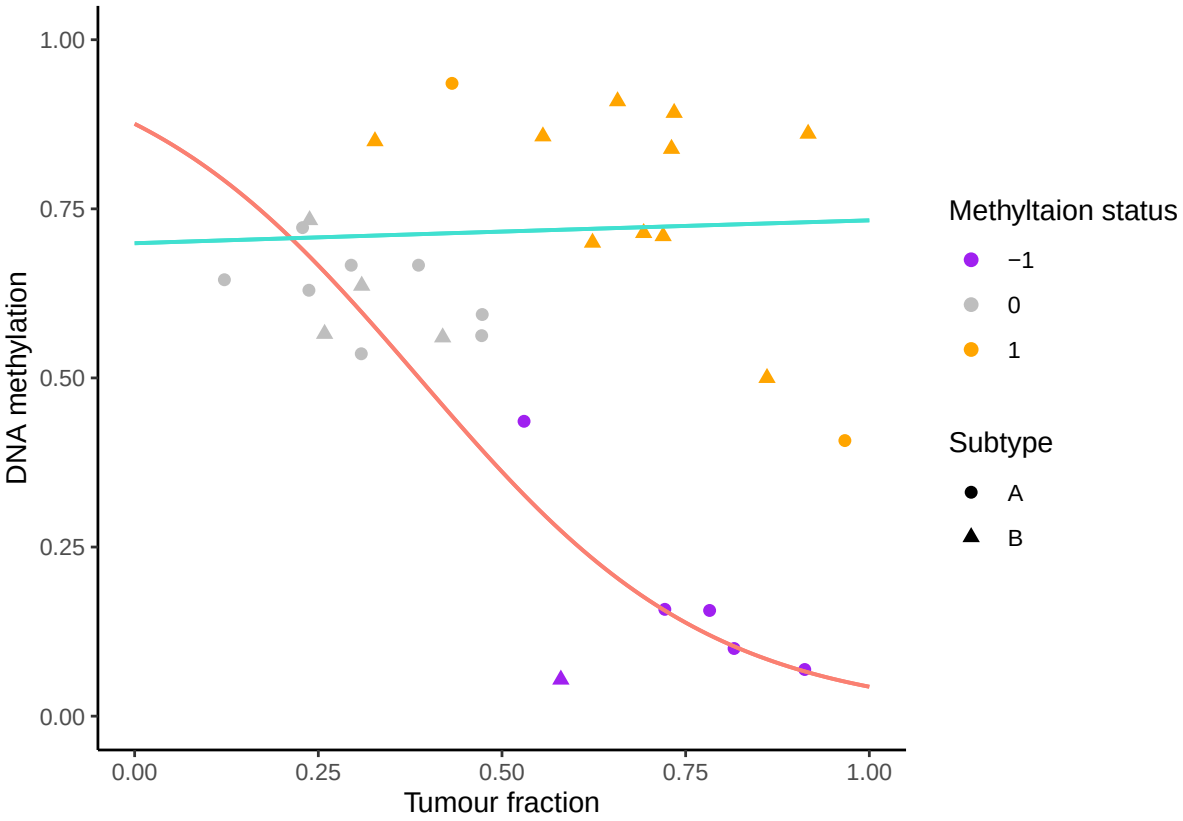
```
##    clus
##     -1  1
##   A  5  2
##   B  1 10
```

Out of 30 samples, the mixture modelling approach classified 18 samples, and most samples are classified correctly.

Visualize the predicted subtypes and their methylation values. Similar as above, shape of the dot corresponds to the ground truth tumour subtype, whereas colour of the dot represents the subtype prediction. The salmon and turquoise lines represents the two mixing components modelled by the algorithm.

```
dt %>%
    cbind(data.table(mixmodel = clus)) %>%
    setnafill(fill = 0, cols = 'mixmodel') %>%
    ggplot(aes(x = tumour_fraction, y = mod / cov, shape = subtype,
               color = factor(mixmodel), group = subtype)) +
    geom_point(size = 2) +
    geom_function(
        linewidth = 0.7, col = 'salmon',
        fun = ~ plot_logit_transformed(.x, parameters(mdl)[, 1])
    ) +
    geom_function(
        linewidth = 0.7, col = 'turquoise',
        fun = ~ plot_logit_transformed(.x, parameters(mdl)[, 2])
    ) +
```

```
    scale_color_manual(values = c('purple', 'grey', 'orange'),
                       name = 'Methyltaion status') +
    scale_shape(name = 'Subtype') +
    expand_limits(x = c(0, 1), y = c(0, 1)) +
    xlab('Tumour fraction') + ylab('DNA methylation')
```



B

Cook's distance implementation

B.1 Cook's distance calculation

This section is fully credited to Dr. George Nicholson who has kindly produced the following mathematical writing.

Locus $i = 1, \ldots, N$, sample $d = 1, \ldots, D$, p-dimensional covariates $\boldsymbol{x}_d$.

$$
\begin{aligned}
Z_{id} &= \arcsin(2Y_{id}/m_{id} - 1) \\
\mathbb{E}[Z_{id}] \mid \boldsymbol{\beta}_i &\approx \boldsymbol{x}_d \boldsymbol{\beta}_i \\
\mathbb{V}(Z_i \mid \boldsymbol{\beta}_i) \equiv \boldsymbol{V}_i &:= \operatorname{diag}_{d=1}^{D} \left(\frac{1 + (m_{id} - 1)\phi_i}{m_{id}} \right)
\end{aligned} \tag{B.1}
$$

The linear model can be expressed:

$$
\boldsymbol{Z}_i \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}_i, \boldsymbol{V}_i) \tag{B.2}
$$

The weighted least squares solution and Cook's distance for Eq. (B.2) can be seen from working with the OLS least squares solution and ordinary definition of Cook's distance to the transformed model Eq. (B.3):

$$
\boldsymbol{V}_i^{-1/2} \boldsymbol{Z}_i \sim \mathcal{N}(\boldsymbol{V}_i^{-1/2} \boldsymbol{X}\boldsymbol{\beta}_i, \boldsymbol{I}) \tag{B.3}
$$

I.e. the OLS solution, hat matrix, fitted values, residuals etc. are:

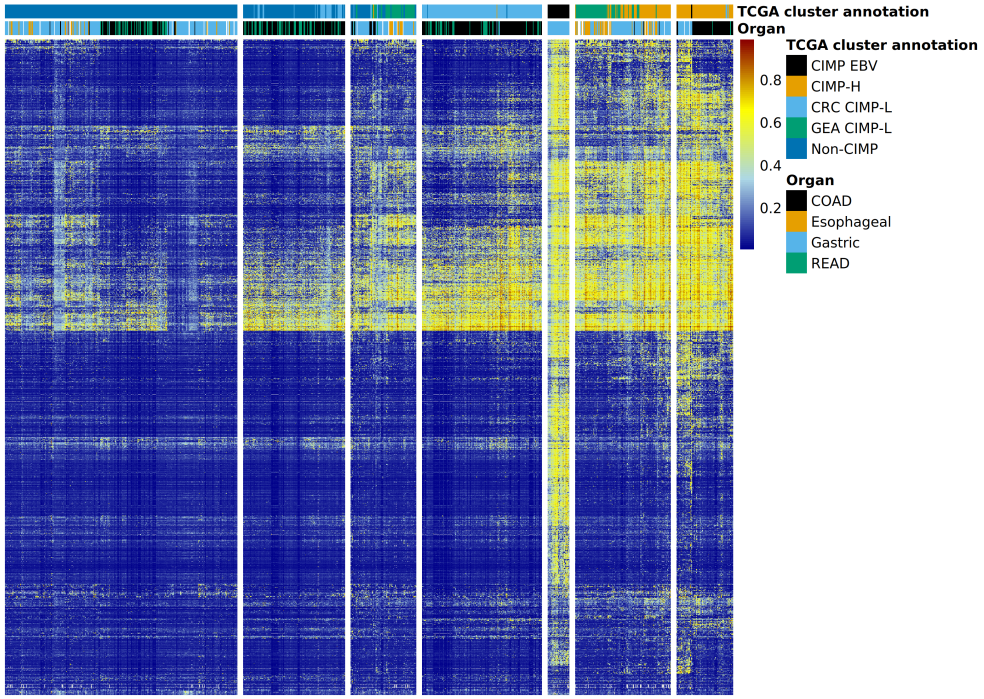
$$\begin{aligned}
\hat{\boldsymbol{\beta}}_i &= (\boldsymbol{X}^T \boldsymbol{V}_i^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{V}_i^{-1} \boldsymbol{Z}_i \\
\boldsymbol{H} &= \boldsymbol{V}_i^{-1/2} \boldsymbol{X} (\boldsymbol{X}^T \boldsymbol{V}_i^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{V}_i^{-1/2} \\
\boldsymbol{V}_i^{-1/2} \hat{\boldsymbol{Z}}_i &= \boldsymbol{V}_i^{-1/2} \boldsymbol{X} \hat{\boldsymbol{\beta}}_i \\
&= \boldsymbol{H} \boldsymbol{V}_i^{-1/2} \boldsymbol{Z}_i \\
\boldsymbol{e}_i &= \boldsymbol{V}_i^{-1/2} \boldsymbol{Z}_i - \boldsymbol{V}_i^{-1/2} \hat{\boldsymbol{Z}}_i \\
s_i^2 &= \frac{\boldsymbol{e}_i^T \boldsymbol{e}_i}{D - p}
\end{aligned} \tag{B.4}$$

The Cook's distance of observation d at locus i is

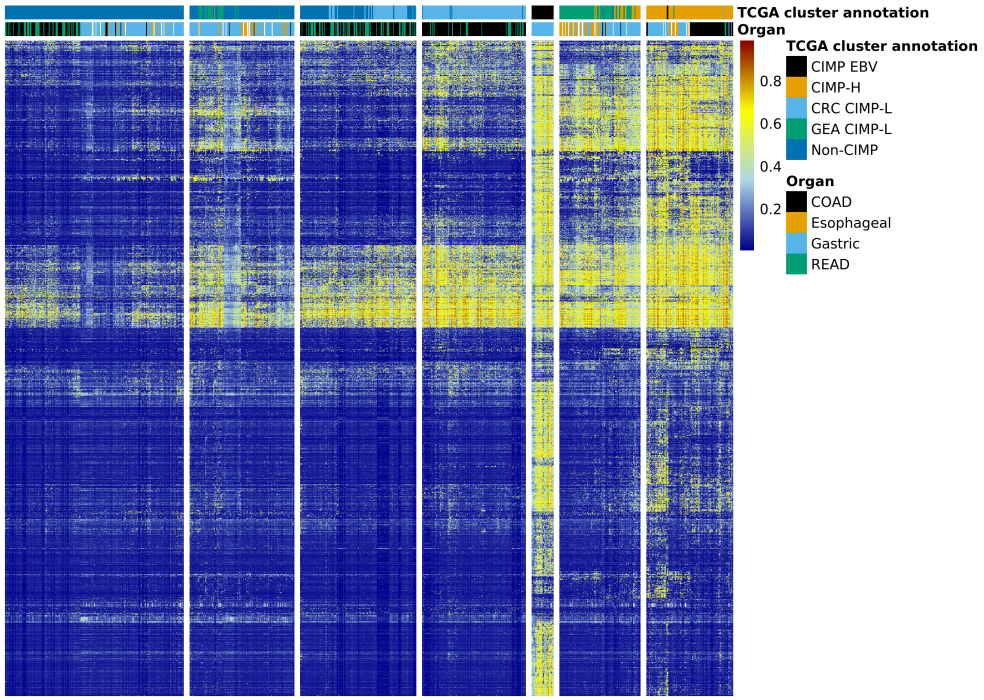
$$\mathcal{C}_{id} := \frac{e_{id}^2}{s_i^2 p} \left[\frac{h_{dd}}{(1 - h_{dd})^2} \right] \tag{B.5}$$

C

Supplementary figures and tables



(a) Recreation of Liu et al. heatmap. Original cluster annotations in Liu et al. are shown in the column metadata, labelled as "TCGA cluster annotation".



(b) Same data but using 0.28 methylation cutoff.

Fig. C.1: Methylation heatmaps of TCGA GI adenocarcinoma data. Data are downloaded from TCGA data portal instead of from the original publication.

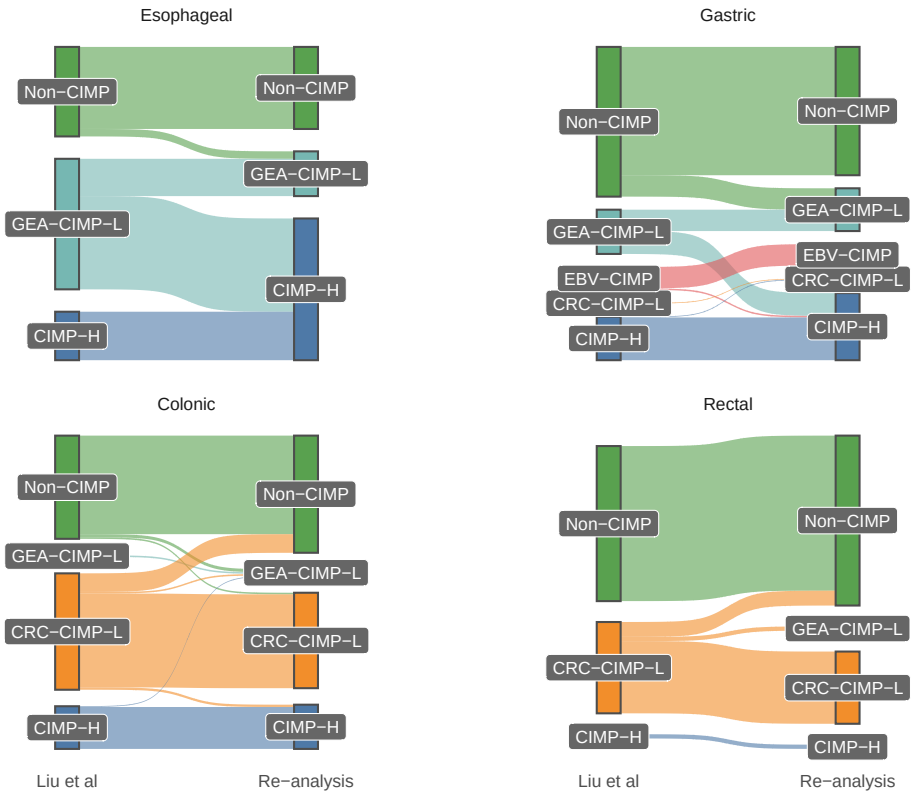


Fig. C.2: Sankey plot comparing published and recreated clusters. Differences in cluster annotations may be due to different data processing pipeline.

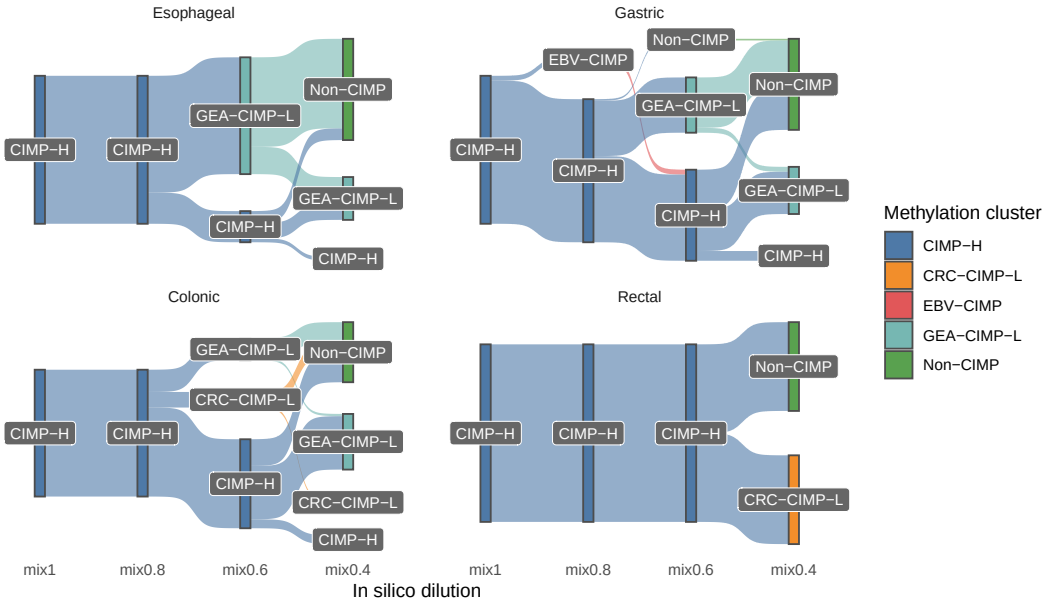


Fig. C.3: Sankey plot to visualize the effect of tumour fraction, facetted by site of origin.

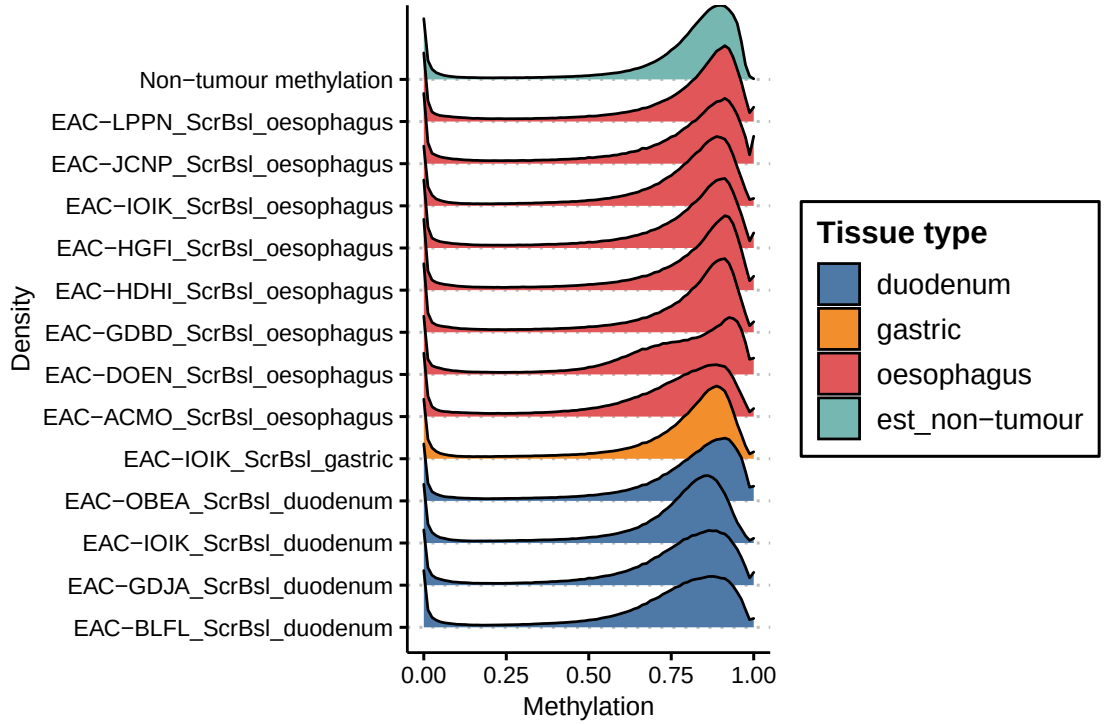


Fig. C.4: Ridgeplot comparing the global methylation landscape of estimated non-tumour methylation and adjacent normal tissues of trial patients. The height of the ridge represents the frequency of CpGs with methylation given on the x-axis. Top row represents the estimated non-tumour methylation, which is very similar to the measured normal tissues. Oesophageal sample from patient EAC-DOEN is contaminated with tumour cells and looks different from other samples.

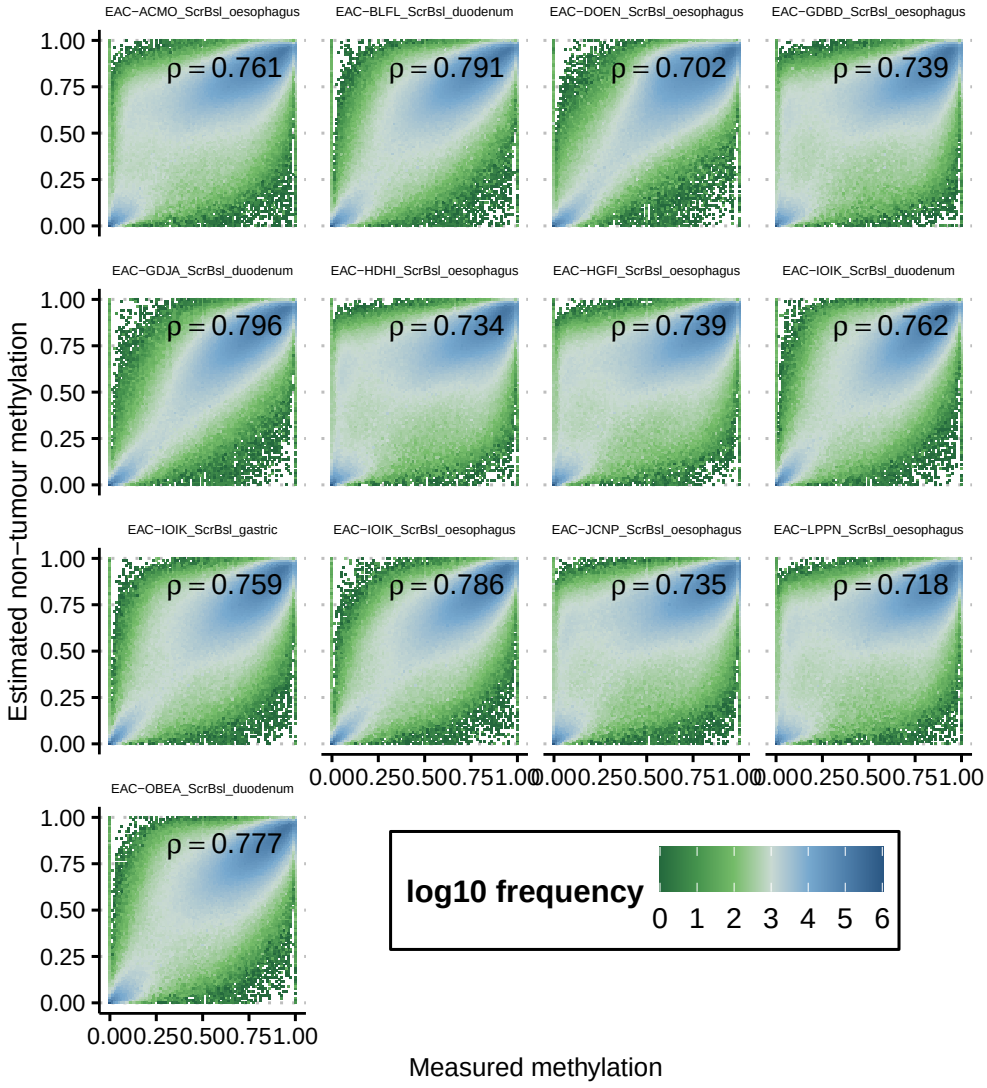


Fig. C.5: 2D density plot with estimated non-tumour methylation on y-axis and the measured sample methylation on x-axis. Spearman's ρ is labelled for each pairwise comparison, p values for the correlation test are all $<2.22e{-}16$.

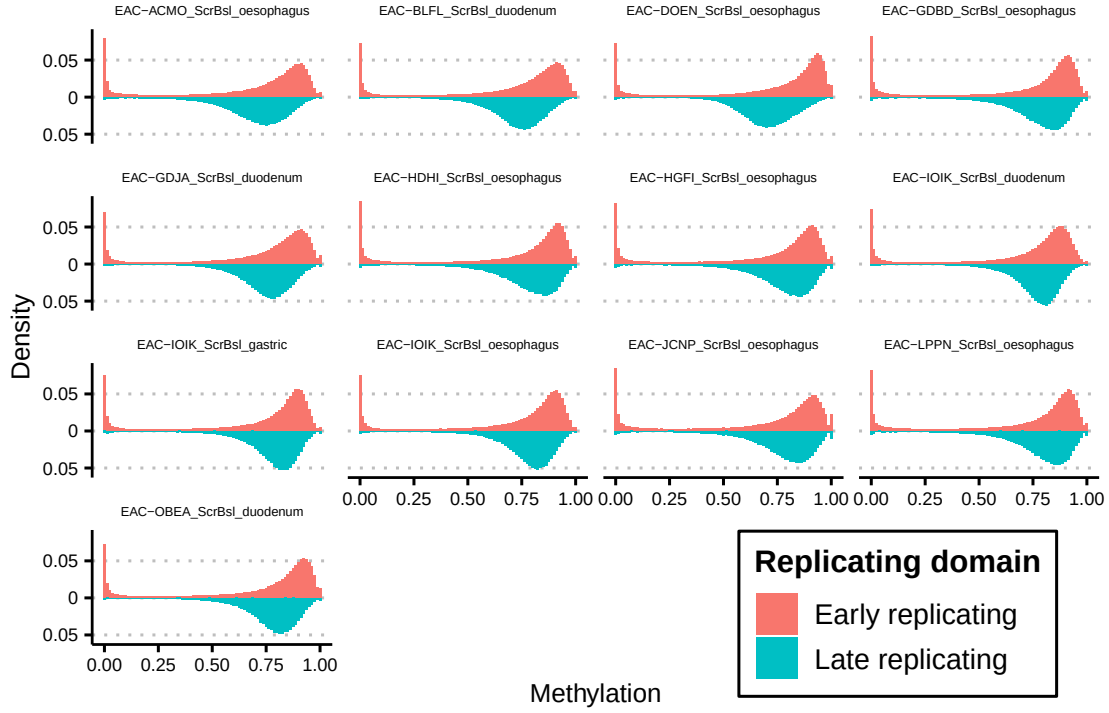


Fig. C.6: Mirrored histograms of replication domain methylation of adjacent normals, with early replicating domain labelled as salmon colour and late replicating domain as turquoise. Y-axis is displayed as density. Note the absence of fully unmethylated CpGs in late replicating domains, which in contrary are present in high proportions in early replicating domains.

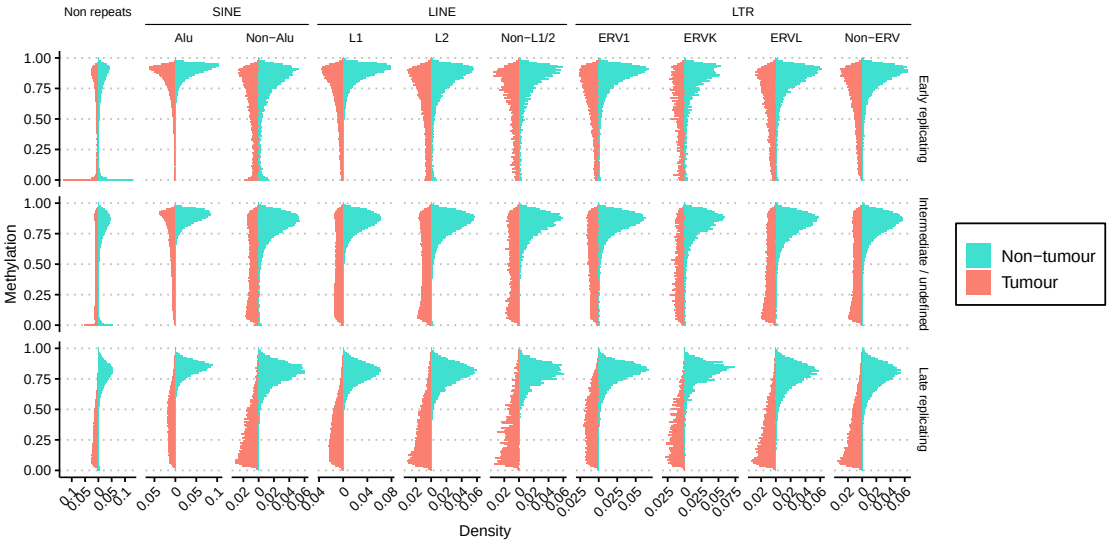
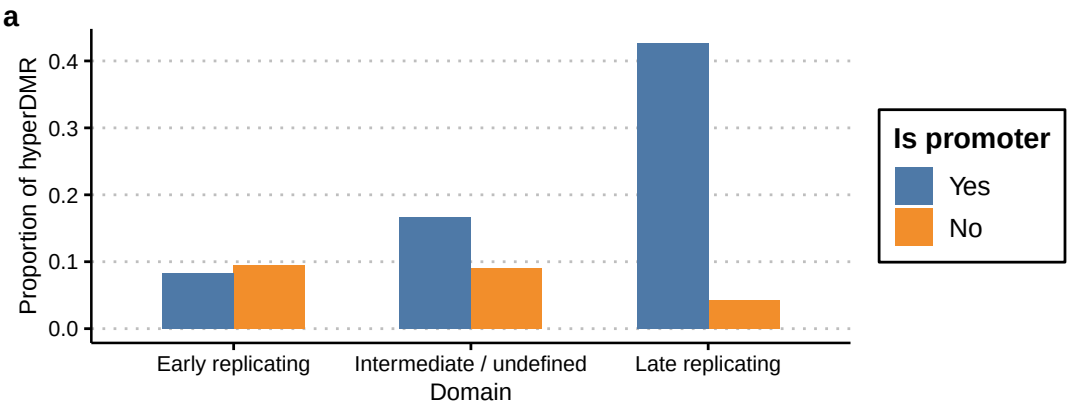


Fig. C.7: Mirrored histograms of RE methylation of tumour and non-tumour, facetted by replication domains, with density on x-axis and methylation on y-axis. Tumour is labelled with salmon colour and non-tumour with turquoise. RE hypomethylation is not apparent in early replicating domain.

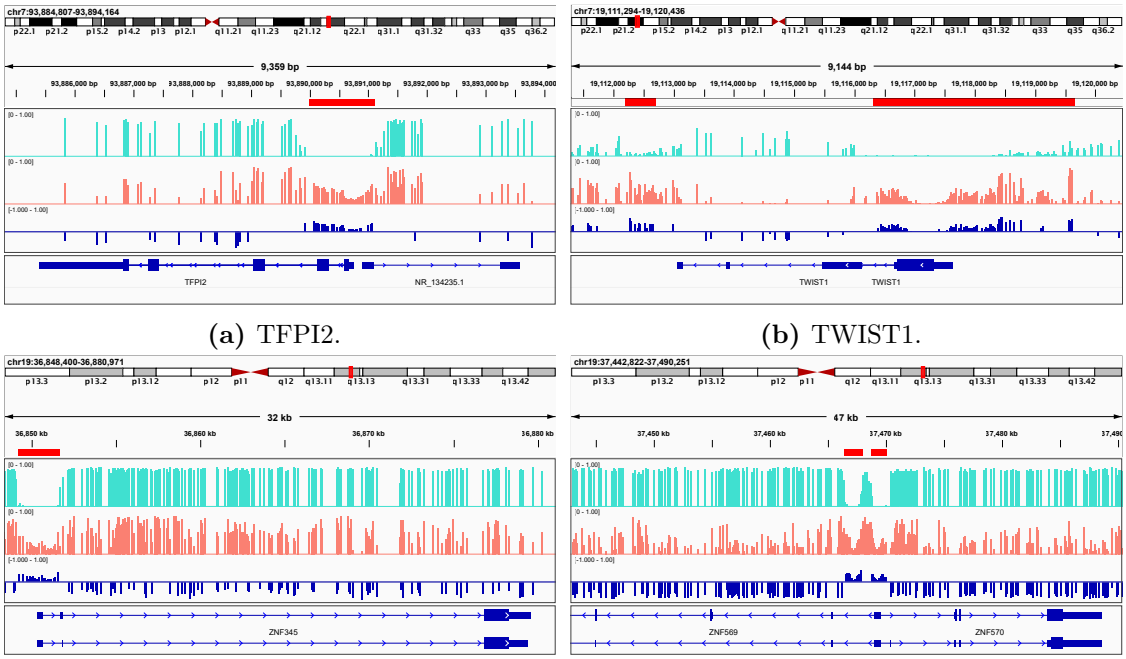
a



b

Is promoter	Domain	hyperDMR	Valid intervals	Odds	Odds ratio	p value
Yes	Early replicating	3258	39195	0.0907	(ref)	
	Late replicating	360	844	0.7440	8.204	4.95e−188
No	Early replicating	556	5882	0.1040	(ref)	
	Late replicating	60	1410	0.0444	0.426	8.69e−10

Fig. C.8: Hypermethylation in late replicating domain is mediated through promoter methylation, defined as bivalent promoter, promoter, or TSS. (**a**) Barchart showing the fraction of hyperDMR that belongs to promoter and non-promoters for each replicating domain. (**b**) Table describing the odds ratio of being hypermethylated in the corresponding group, with early replicating domain as reference. Binomial regression is used to obtain test statistics. Promoters in late replicating domain are strongly enriched for hyperDMR compared to early replicating domain, while non-promoters are depleted for hyperDMR.



(a) TFPI2. (b) TWIST1.

(c) ZNF345. (d) ZNF569.

Fig. C.9: Validation of BO-specific methylation panel defined by Chettouh et al. at single-base resolution. Top track represents the modelled non-tumour methylation (turquoise), middle track represents the tumour methylation (salmon), and lower track represents significant DMR (blue). DMR ranges from -1 to 1, with positive values being more methylated in tumour. HyperDMRs are manually annotated with red bar at top.

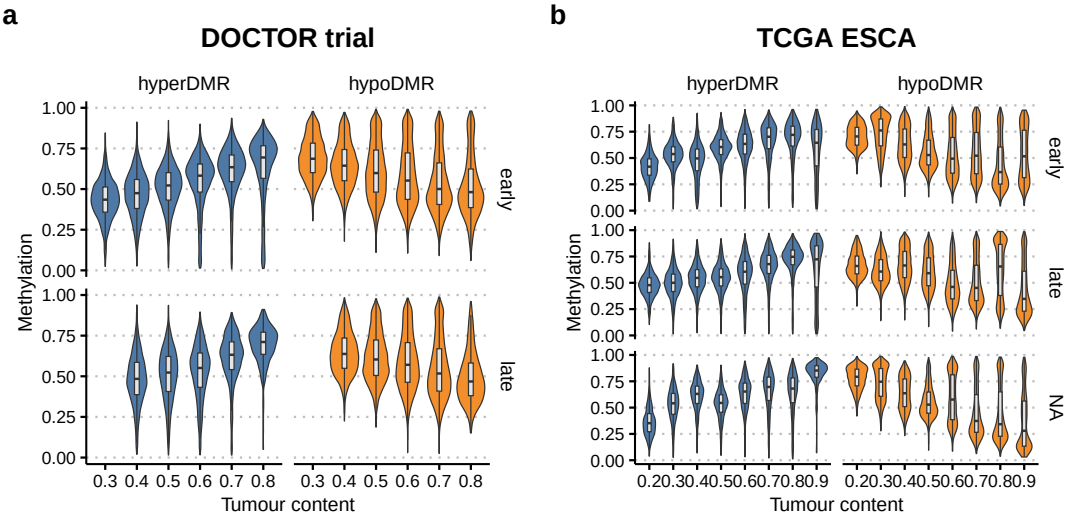


Fig. C.10: OAC-specific methylation marks facetted by stage. (**a**) Violin plot of samples from the DOCTOR trial and (**b**) TCGA ESCA cohort. There is a trend between methylation and tumour purity in both early stage (I and II) and late stage (III and IV) disease.

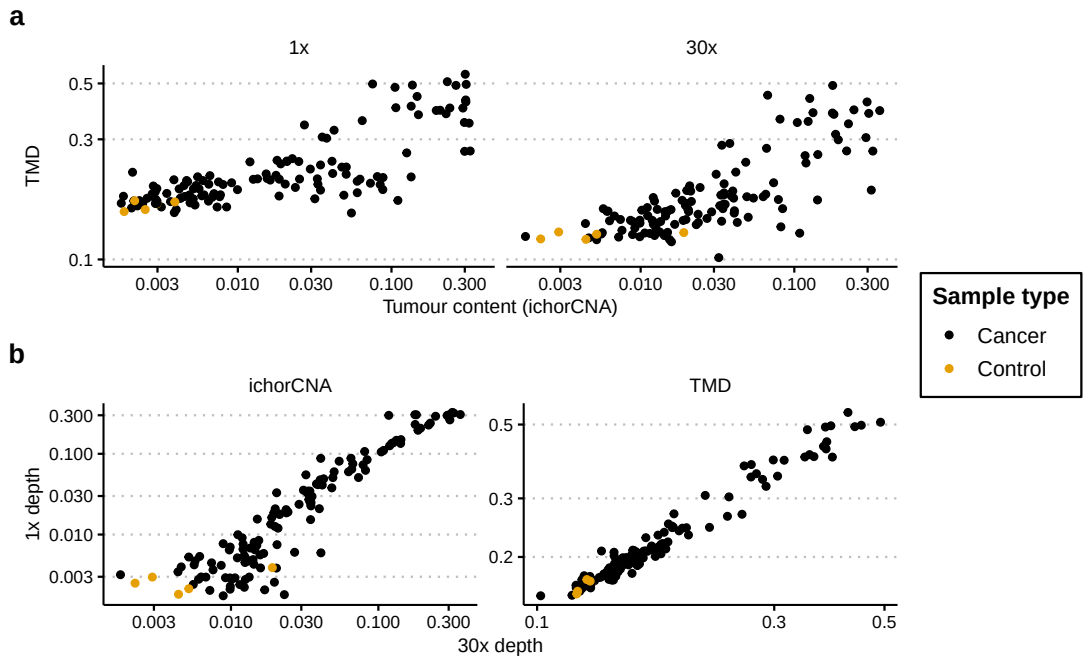


Fig. C.11: Comparison of *ichorCNA* and TMD estimates at different sequencing depth of cfDNA samples. Plot is on log-log scale for better visualization of low tumour content samples. (**a**) Comparing TMD against *ichorCNA* for each sequencing depth. (**b**) Comparing 1x against 30x sequencing depth for each method.

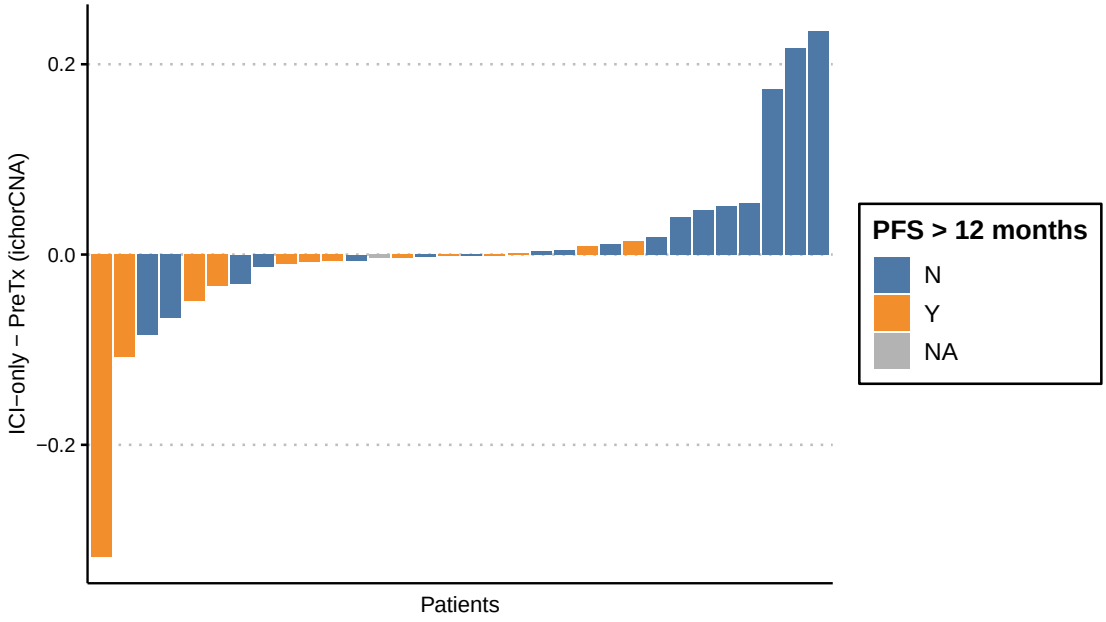


Fig. C.12: Barplot of change in tumour content at ICI-only compared to baseline by *ichorCNA*. Samples are ordered from negative to positive change, blue indicates NCB and orange indicates CB.

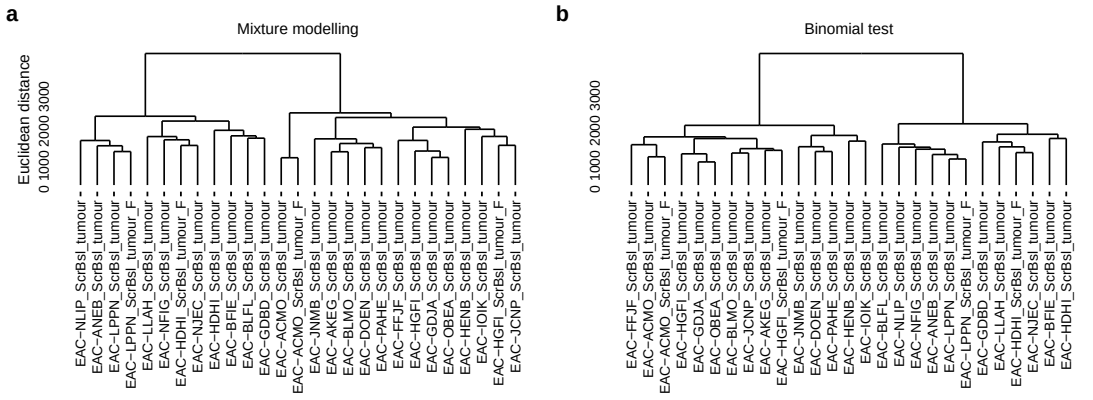


Fig. C.13: Dendrograms of the hierarchical clustering of methylation subtype.

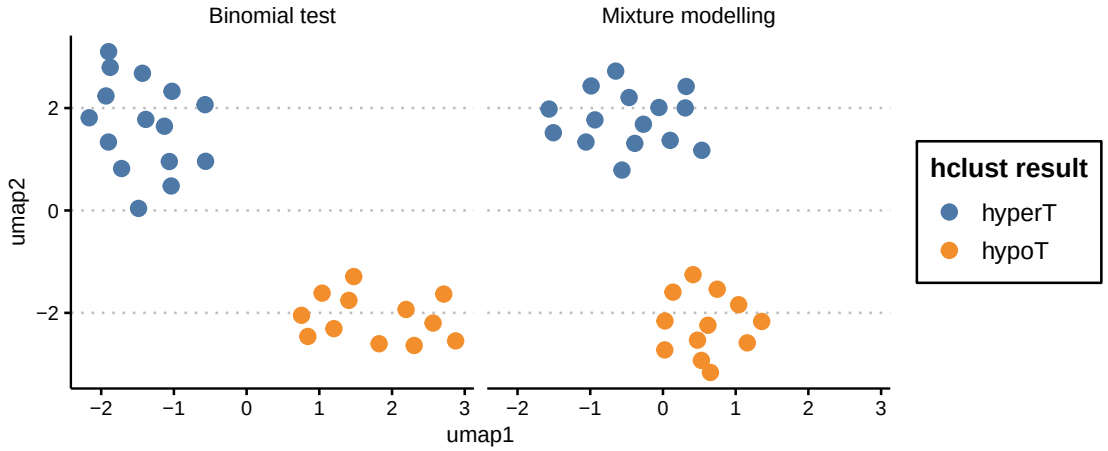


Fig. C.14: UMAPs on the PCA-transformed, genome-wide tumour DNA fraction corrected methylation status.

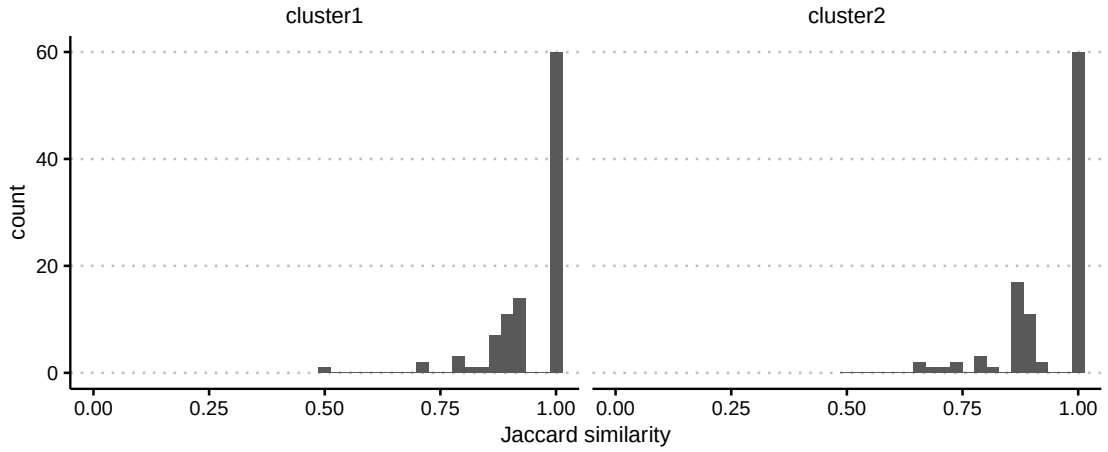


Fig. C.15: Jaccard similarity of methylation clusters upon bootstrapping 100 times using the *fpc* package.

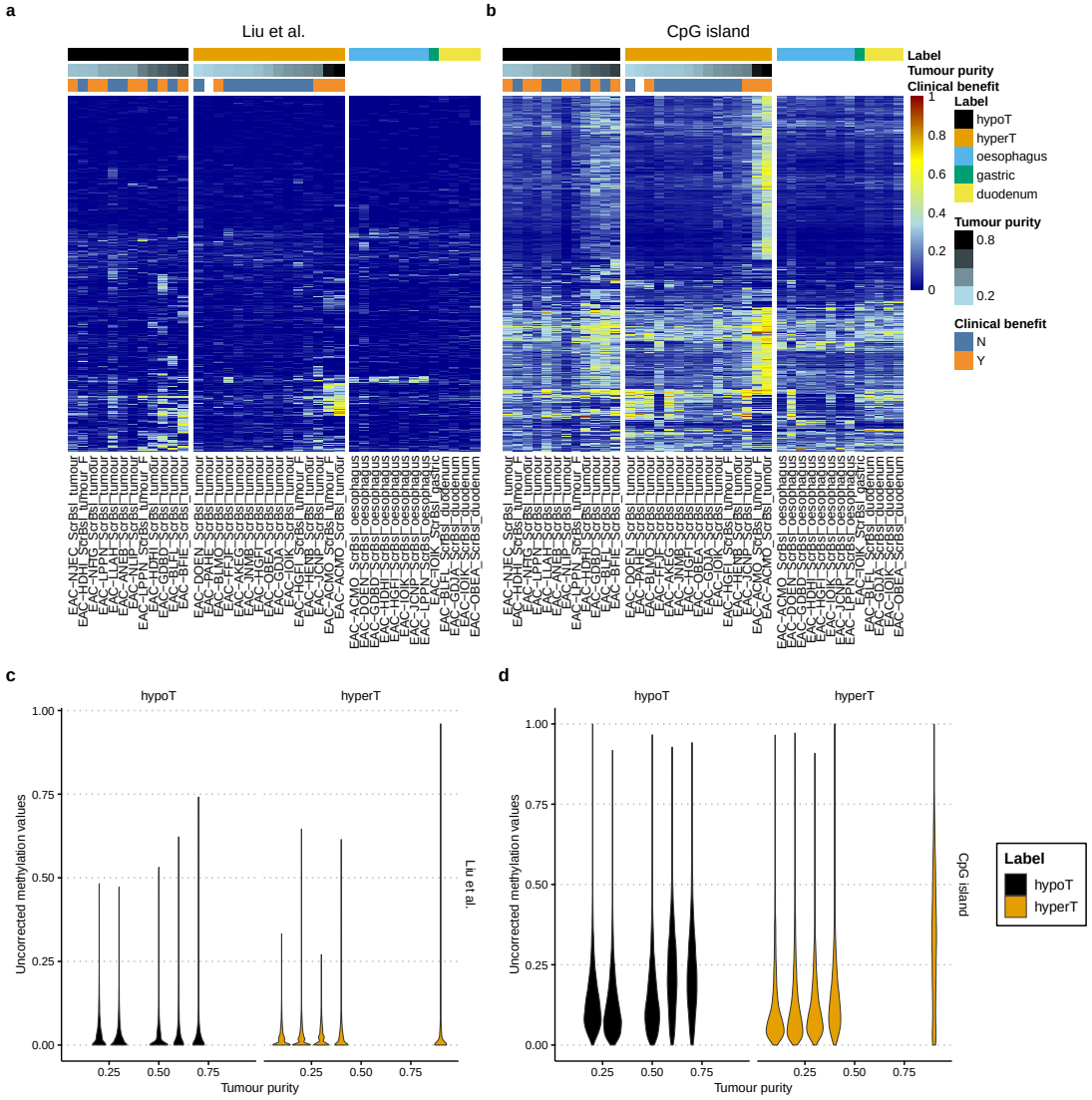


Fig. C.16: Visualization of methylation values of tumour subtypes at potential CIMP loci. (**a,b**) Heatmap of uncorrected methylation values, with columns representing samples, grouped by methylation subtypes and tumour purity. Loci in (**a**) are chosen as defined in Liu et al., and loci in (**b**) are filtered CpG islands with < 0.3 average methylation in normal samples, and > 0.3 methylation in at least 1 tumour sample. (**c,d**) Violin plots of uncorrected methylation values of hypoT and hyperT clusters at the potential CIMP loci defined in (**c**) Liu et al. and (**d**) filtered CpG islands.

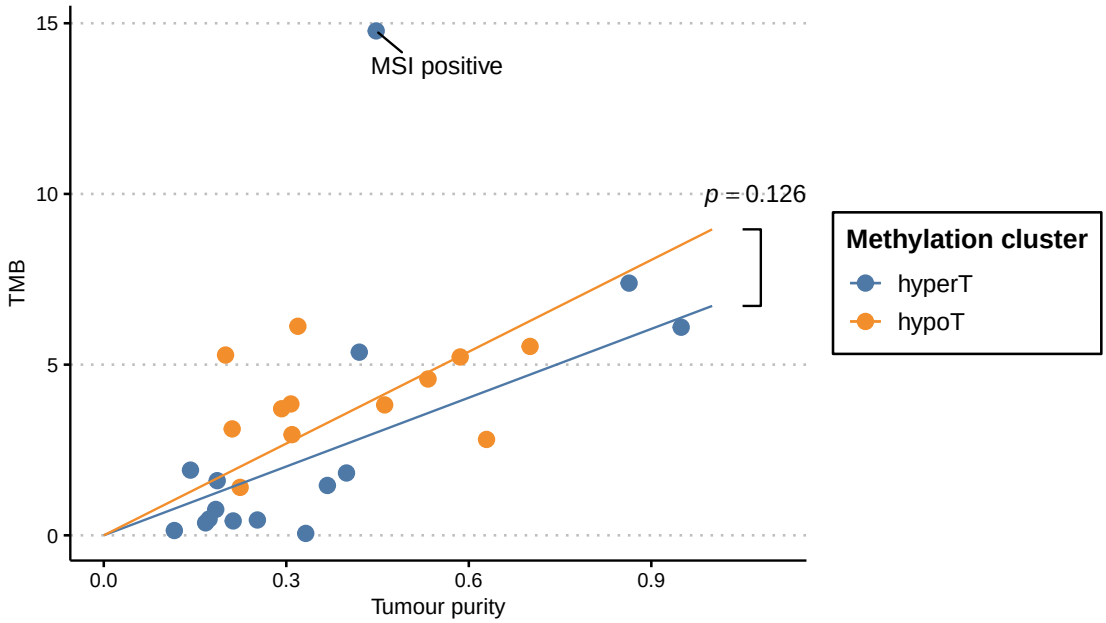


Fig. C.17: TMB in subtype with linear regression against tumour purity. The MSI sample is excluded from the model for being an outlier. Model design forced a 0 intercept and used the interaction between subtype and tumour purity as covariate. Model contrast is performed between the two model coefficients at 100% tumour purity and $p = 0.126$.

a

SBS	Slope	p value	FDR
SBS1	−0.31	0.02	0.11
SBS2	−0.01	0.57	0.75
SBS3	0.02	0.30	0.66
SBS5	−0.09	0.45	0.75
SBS8	0.05	0.00	0.06
SBS17a	0.11	0.06	0.19
SBS17b	0.23	0.02	0.11
SBS18	−0.02	0.67	0.79
SBS28	0.00	NA	NA
SBS30	0.00	0.14	0.37
SBS35	−0.01	0.48	0.75
SBS40	0.02	0.79	0.79
SBS41	−0.01	0.56	0.75
SBS44	0.02	0.73	0.79

b

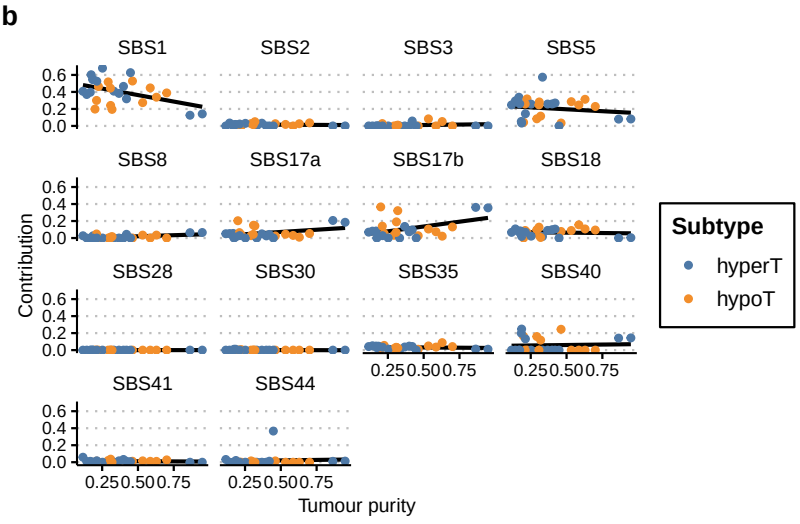


Fig. C.18: (**a**) Results of linear regression including slope coefficients, p values, and fdr values. Prior to FDR correction, significant associations between mutational signatures and tumour purity are found upon linear regression for SBS1, SBS8, and SBS17b. No significant hits remain after controlling for multiple testing. (**b**) Scatterplot of mutational signatures against tumour purity. Colours represent the methylation subtypes. Black lines represent the fitted linear model.

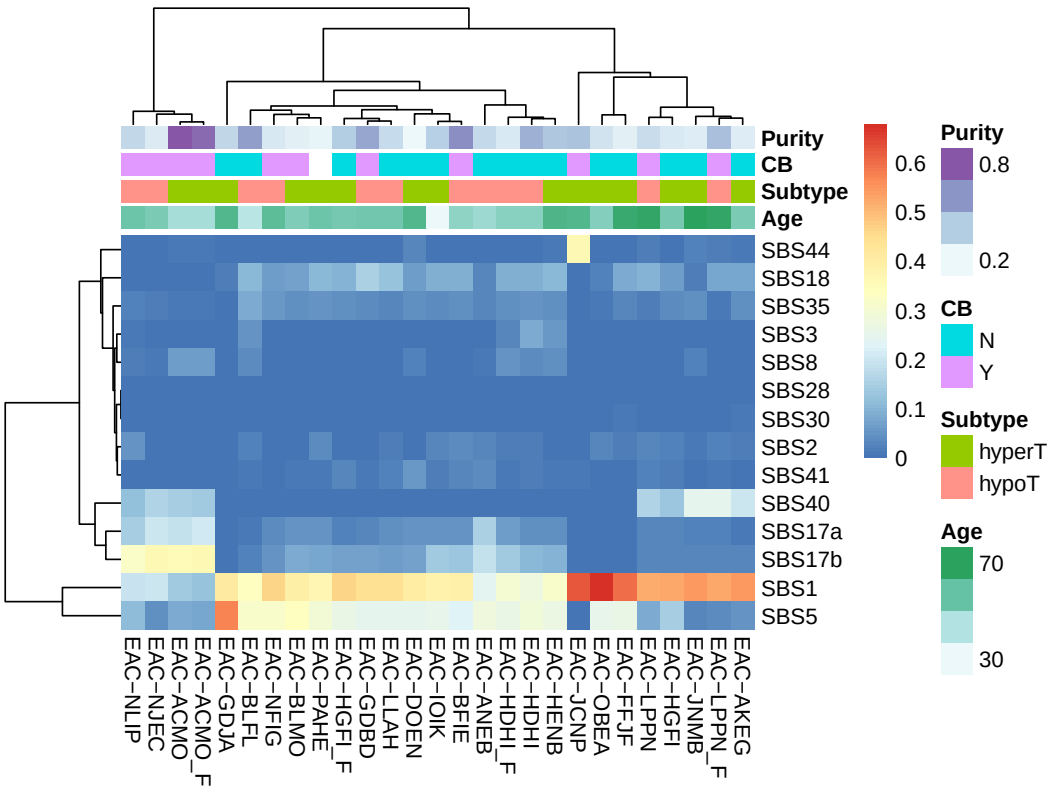


Fig. C.19: Heatmap of contributions of 14 pre-defined mutational signatures. Each column represents one sample, and each row represents one signature. As expected, the sample with MSI has high SBS44, which is a result of mismatch repair (MMR) deficiency.

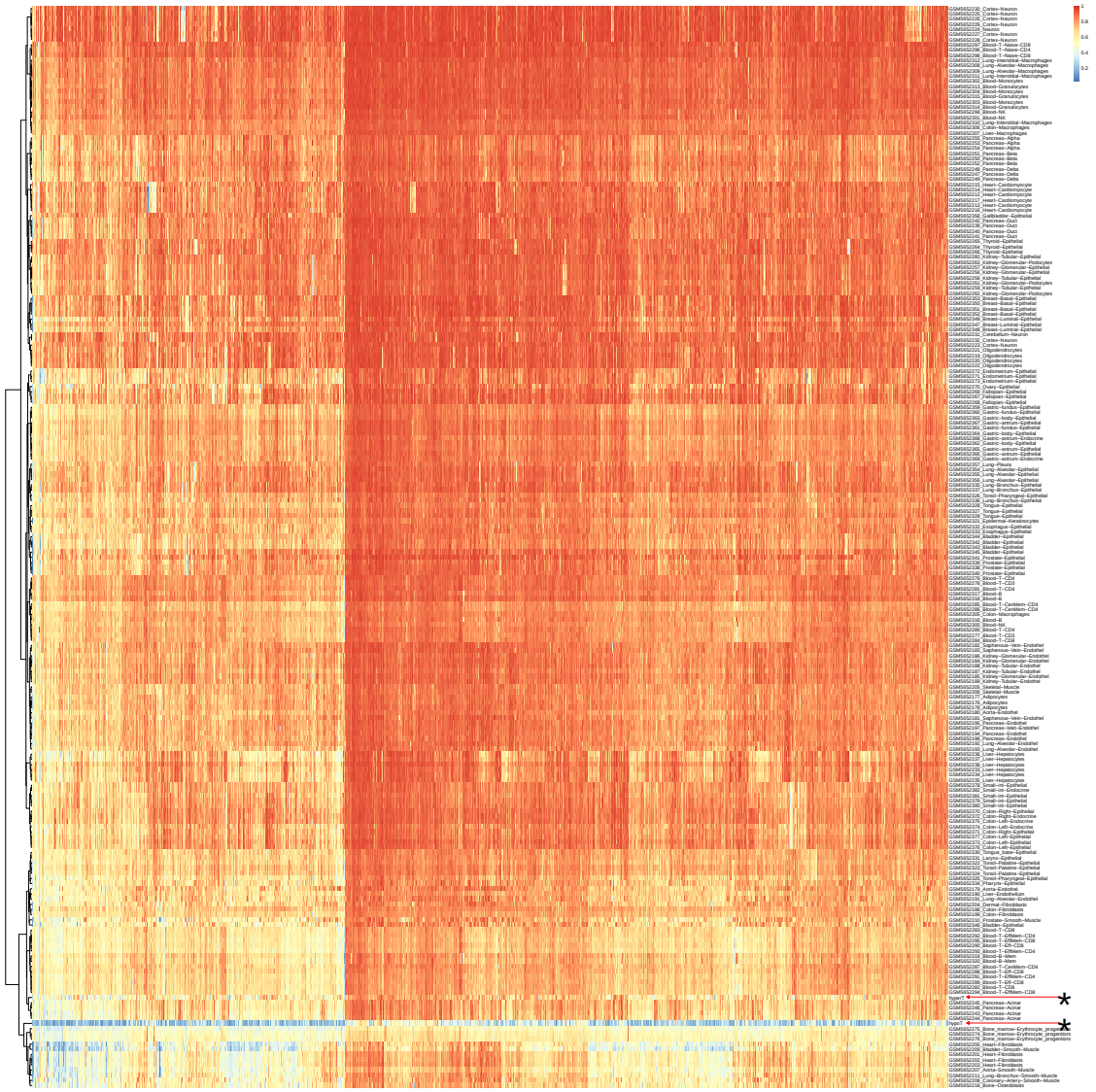


Fig. C.20: Heatmap of tumour subtype with normal cell types. The position of hypoT and hyperT subtypes are marked with red arrow with an asterisk. HyperT is closest to pancreatic acinar cells, whereas hypoT is closest to a group of mesenchymal cells including fibroblast, smooth muscle, erythrocyte progenitors, and osteoblasts.

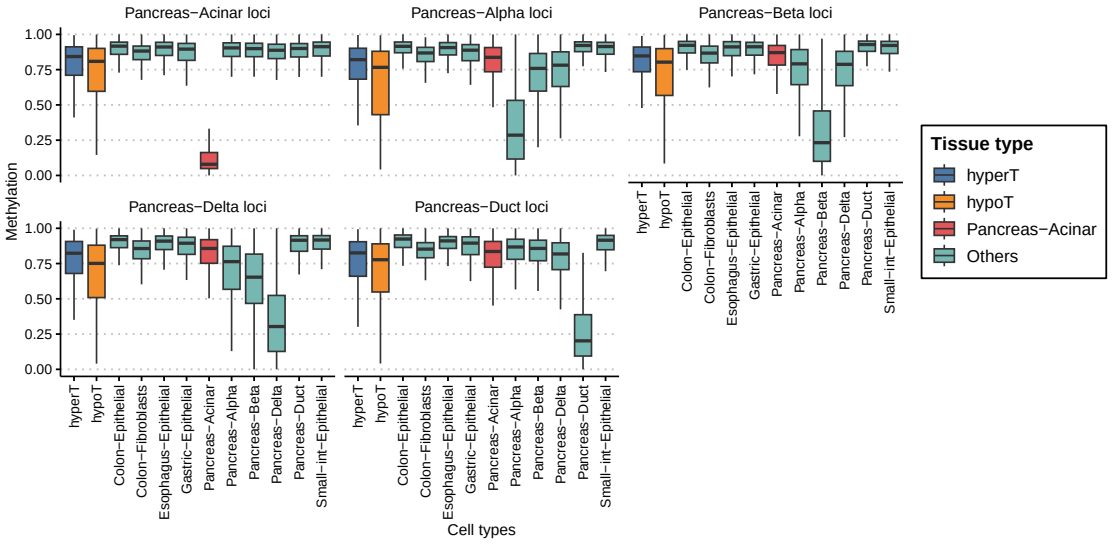


Fig. C.21: Boxplot of methylation of tumour subtypes and other gastrointestinal tissues at pancreas-specific loci. 5 sets of loci are chosen, which are specific for different cell types in the pancreas, presented in the correspondingly named facets. Colours represent tissue types, including hypoT and hyperT subtypes, pancreatic acinar cells, and others. Data of normal cells and loci are taken from Loyfer et al. The normal cells are hypomethylated in their corresponding specific loci, and are methylated otherwise. HyperT and hypoT remains methylated for all pancreas-specific loci.

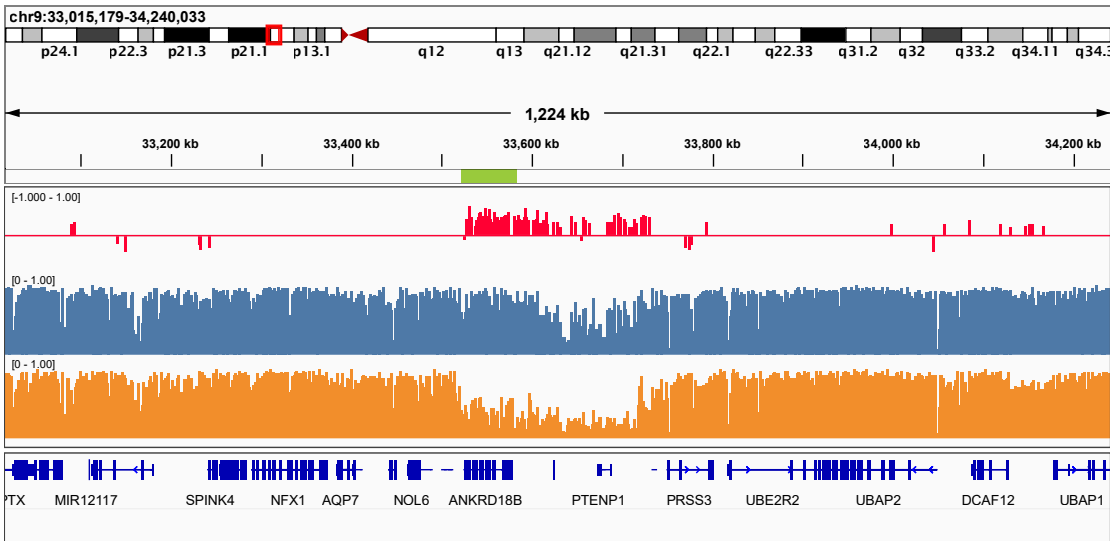


Fig. C.22: Zoomed out view of the methylation landscape of tumour subtypes at ANKRD18B. The location of the gene is marked in green above the DMR track.

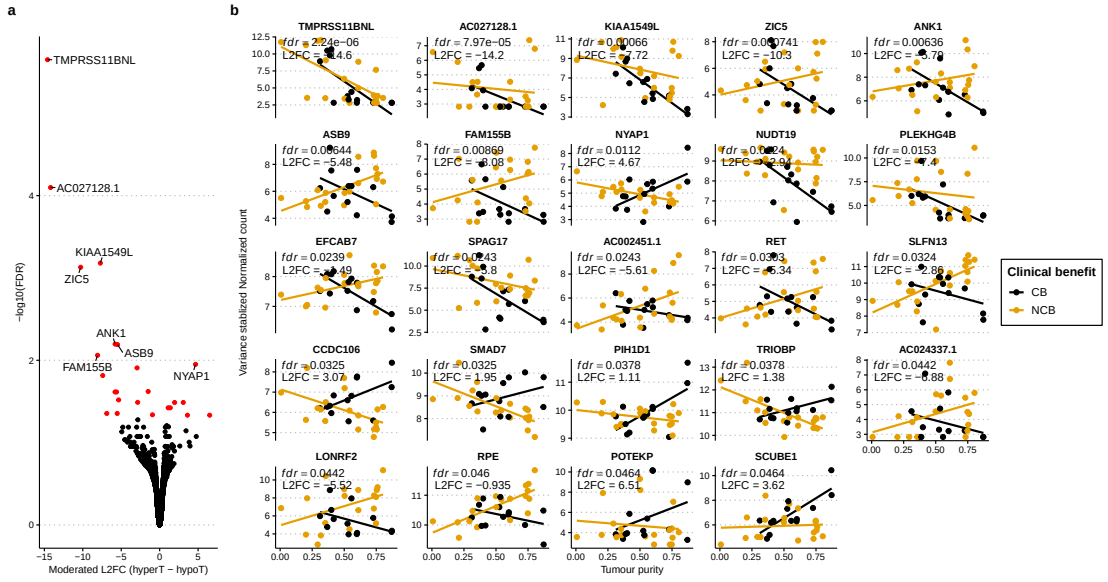


Fig. C.23: (**a**) Volcano plot of DEG testing of CB against NCB in the tumour compartment. Red points indicate $fdr < 0.05$. (**b**) Scatterplot of the variance stabilized gene counts against estimated tumour purity. A linear regression line is fitted for each subtype for visualization purpose only.

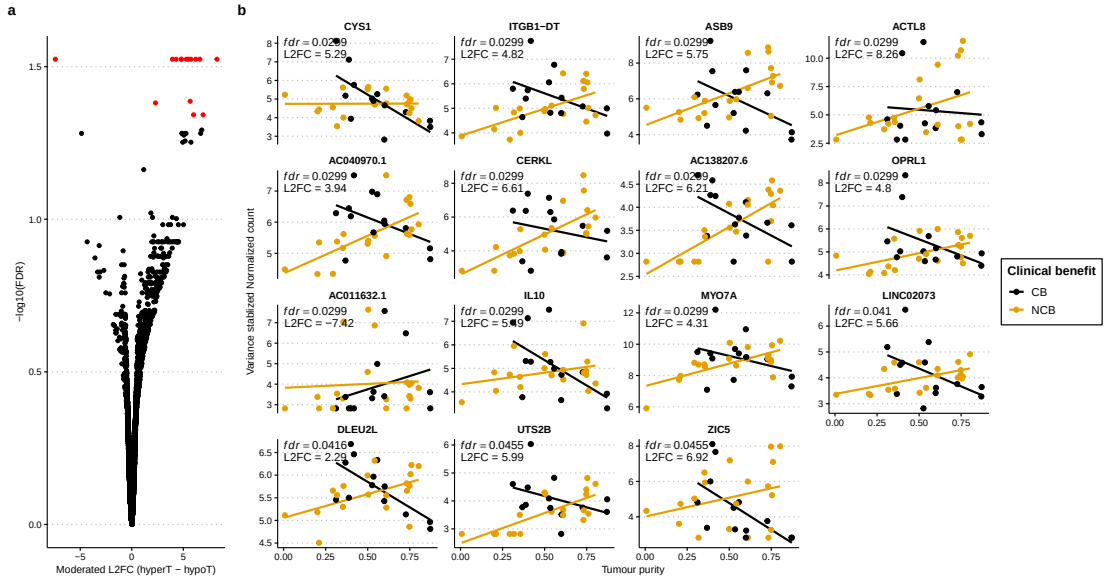


Fig. C.24: (**a**) Volcano plot of DEG testing of CB against NCB in the non-tumour compartment. Red points indicate $fdr < 0.05$. (**b**) Scatterplot of the variance stabilized gene counts against estimated tumour purity. A linear regression line is fitted for each subtype for visualization purpose only.

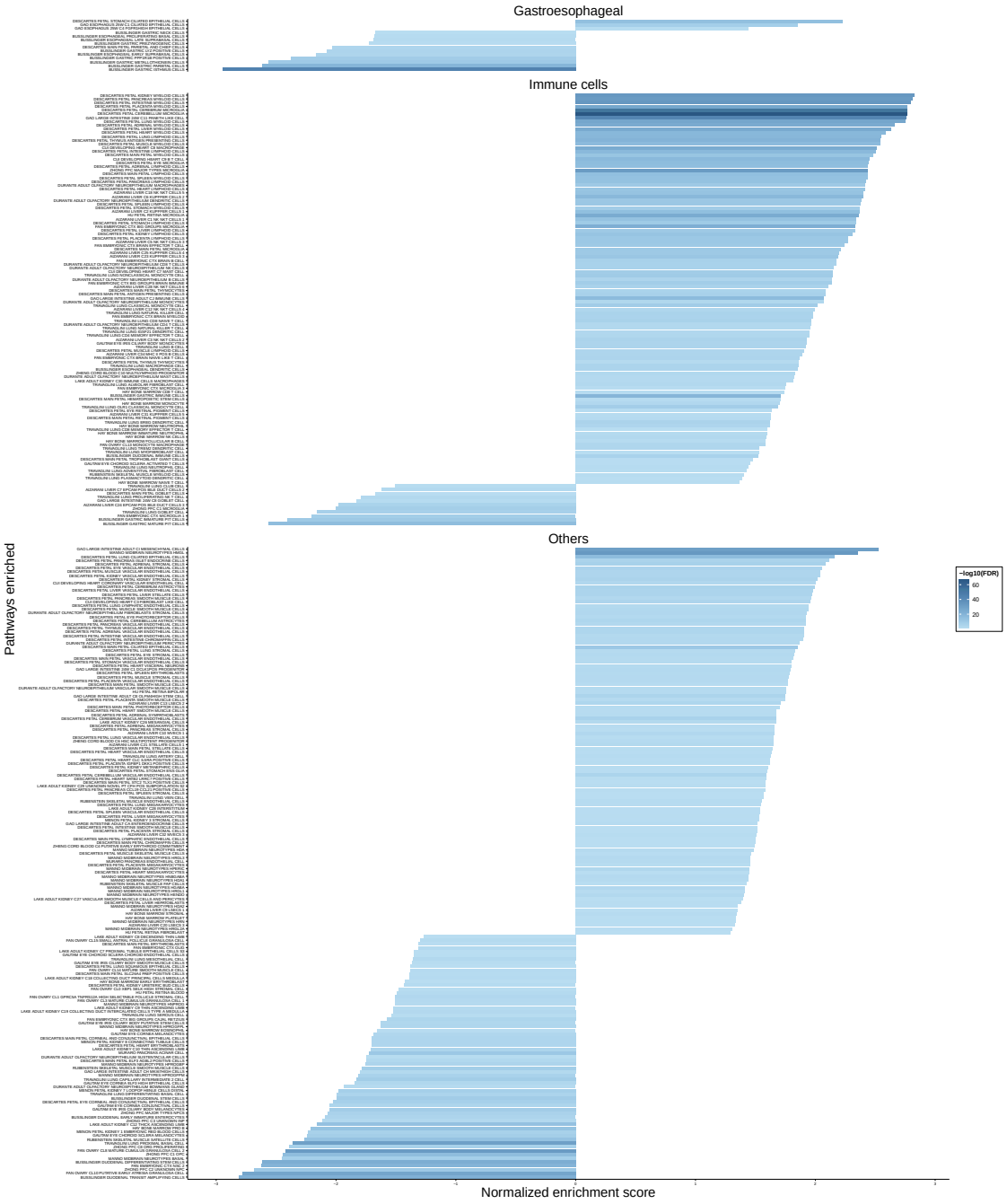


Fig. C.25: GSEA on DEG of CB versus NCB in the non-tumour compartment using the C8 gene set from MSigDB. Only pathways with $fdr < 0.05$ are shown. Positive score suggest relative enrichment in CB samples, and negative score suggest relative enrichment in NCB samples. 104 of 115 significant immune cell signatures (90.4%) are enriched in CB. 11 out of 14 significant gastroesophageal cell signatures (78.6%) are enriched in NCB.

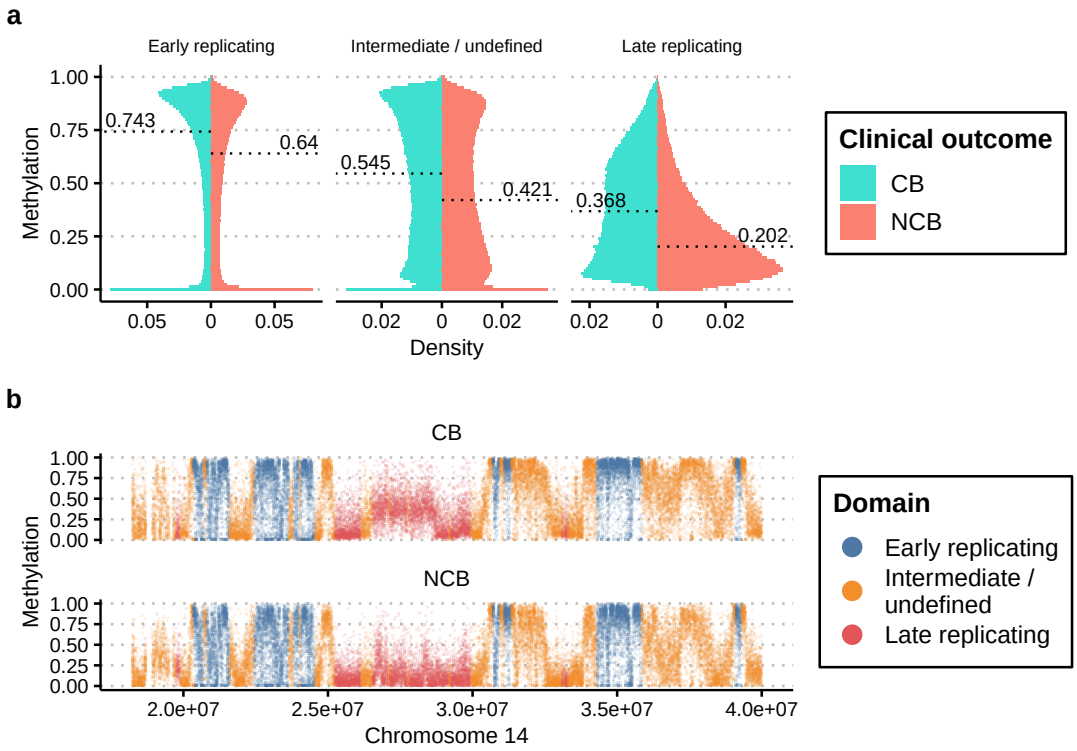


Fig. C.26: Genome-wide methylation landscapes of CB and NCB tumours. (**a**) Mirrored histogram of methylation in different repD. (**b**) Methylation landscapes with sequence context.

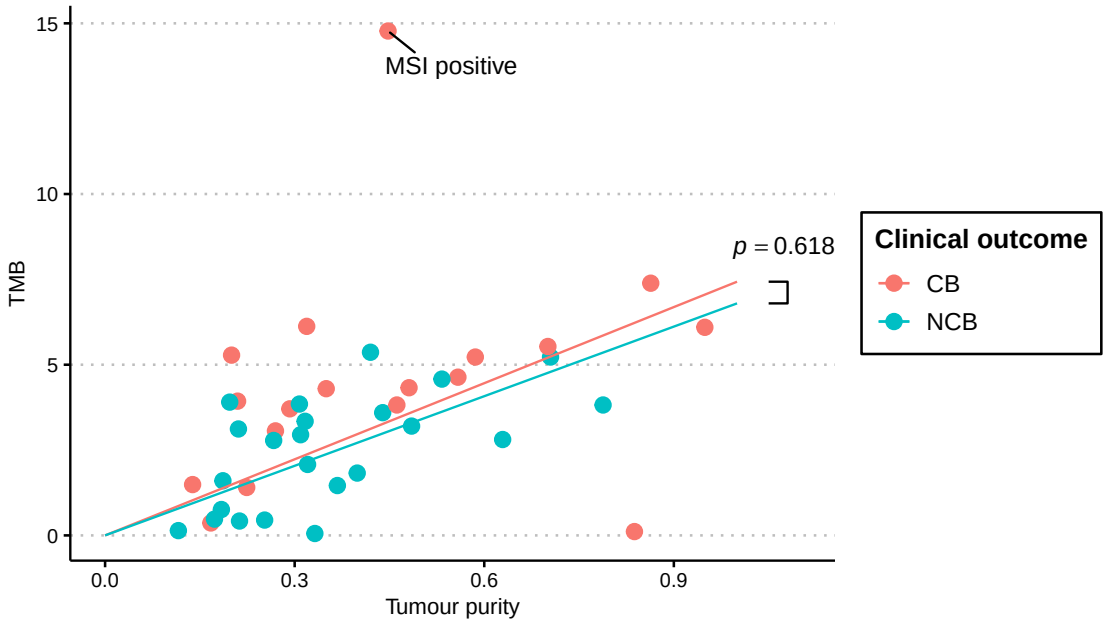


Fig. C.27: Linear regression of TMB against tumour purity in CB versus NCB. The MSI sample is excluded from the model for being an outlier. Model design forced a 0 intercept and used the interaction between clinical outcome and tumour purity as covariate. Model contrast is performed between the two model coefficients at 100% tumour purity and $p = 0.618$.

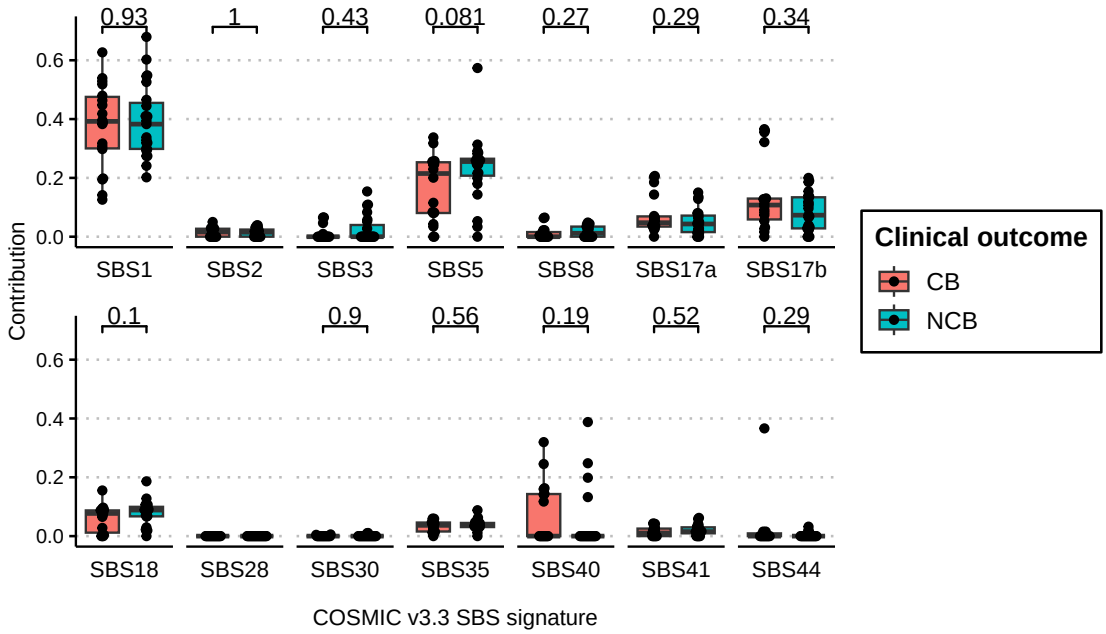


Fig. C.28: Boxplot of mutational signatures in CB versus NCB. Statistics is performed using Wilcoxon test. No statistically significant association is found between clinical outcome and SBS signatures.

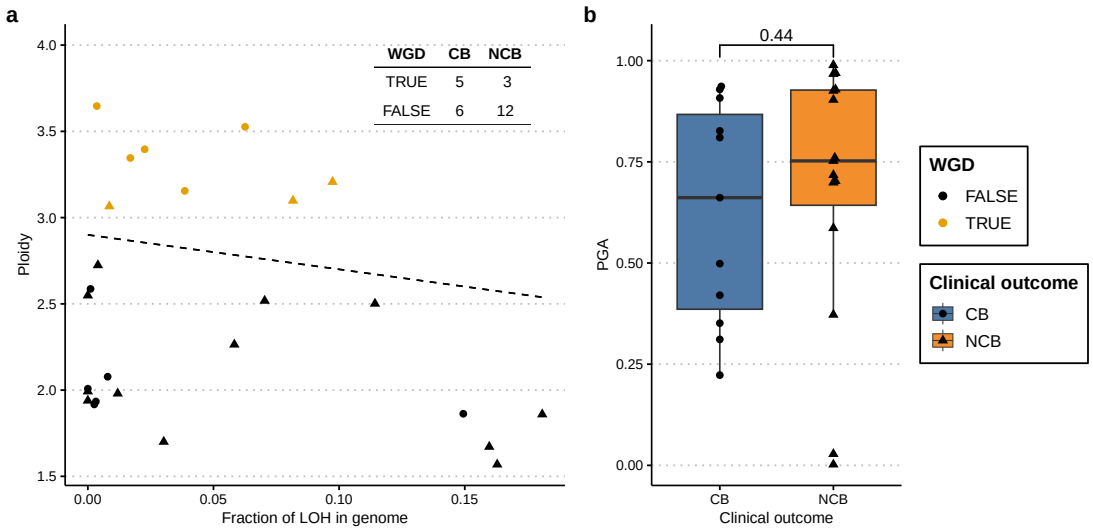


Fig. C.29: Genome instability in CB versus NCB. (**a**) Ploidy of each sample against LOH fraction. Black line represents the cutoff used for identifying WGD. Contingency table at top right corner shows that WGD has no significant evidence of relationship with clinical outcome (Fisher's test, $p = 0.2183$). (**b**) Boxplot of PGA in CB versus NCB. Statistics is performed using Wilcoxon test and is not significant.

References

[1] Douglas Hanahan. 'Hallmarks of Cancer: New Dimensions'. In: *Cancer Discovery* 12.1 (Jan. 2022), pp. 31–46. (Visited on 11/07/2023).

[2] M. T. Barrett et al. 'Evolution of Neoplastic Cell Lineages in Barrett Oesophagus'. In: *Nature Genetics* 22.1 (May 1999), pp. 106–109.

[3] Eric Smith et al. 'Similarity of Aberrant DNA Methylation in Barrett's Esophagus and Esophageal Adenocarcinoma'. In: *Molecular Cancer* 7.1 (Oct. 2008), p. 75. (Visited on 23/09/2023).

[4] Hector Alvarez et al. 'Widespread Hypomethylation Occurs Early and Synergizes with Gene Amplification during Esophageal Carcinogenesis'. In: *PLOS Genetics* 7.3 (Mar. 2011), e1001356. (Visited on 23/09/2023).

[5] Andrew M. Kaz et al. 'DNA Methylation Profiling in Barrett's Esophagus and Esophageal Adenocarcinoma Reveals Unique Methylation Signatures and Molecular Subclasses'. In: *Epigenetics* 6.12 (Dec. 2011), pp. 1403–1412. (Visited on 23/09/2023).

[6] Enping Xu et al. 'Genome-Wide Methylation Analysis Shows Similar Patterns in Barrett's Esophagus and Esophageal Adenocarcinoma'. In: *Carcinogenesis* 34.12 (Dec. 2013), pp. 2750–2756. (Visited on 23/09/2023).

[7] Hamza Chettouh et al. 'Methylation Panel Is a Diagnostic Biomarker for Barrett's Oesophagus in Endoscopic Biopsies and Non-Endoscopic Cytology Specimens'. In: *Gut* 67.11 (Nov. 2018), pp. 1942–1949. (Visited on 23/09/2023).

[8] SriGanesh Jammula et al. 'Identification of Subtypes of Barrett's Esophagus and Esophageal Adenocarcinoma Based on DNA Methylation Profiles and Integration of Transcriptome and Genome Data'. In: *Gastroenterology* 158.6 (May 2020), 1682–1697.e1. (Visited on 22/06/2023).

[9] *Immune Checkpoint Inhibitors - NCI*. https://www.cancer.gov/about-cancer/treatment/types/immunotherapy/checkpoint-inhibitors. cgvArticle. 9/24/2019 - 08:00. (Visited on 23/09/2023).

[10] Shimpei Eguchi et al. 'Durable Response after Discontinuation of Pembrolizumab Therapy for Intrahepatic Cholangiocarcinoma: A Case Report'. In: *Clinical Journal of Gastroenterology* 14.3 (June 2021), pp. 858–865. (Visited on 23/09/2023).

[11] Center for Drug Evaluation and Research. 'FDA Approves Pembrolizumab for Esophageal or GEJ Carcinoma'. In: *FDA* (Mon, 03/22/2021 - 17:45). (Visited on 08/09/2023).

[12] S. Y. Rha et al. 'VP1-2023: Pembrolizumab (Pembro) plus Chemotherapy (Chemo) as First-Line Therapy for Advanced HER2-negative Gastric or Gastroesophageal Junction (G/GEJ) Cancer: Phase III KEYNOTE-859 Study'. In: *Annals of Oncology* 34.3 (Mar. 2023), pp. 319–320. (Visited on 24/09/2023).

[13] Kohei Shitara et al. 'Nivolumab plus Chemotherapy or Ipilimumab in Gastro-Oesophageal Cancer'. In: *Nature* 603.7903 (Mar. 2022), pp. 942–948.

[14] Yelena Y. Janjigian et al. 'First-Line Nivolumab plus Chemotherapy versus Chemotherapy Alone for Advanced Gastric, Gastro-Oesophageal Junction, and Oesophageal Adenocarcinoma (CheckMate 649): A Randomised, Open-Label, Phase 3 Trial'. In: *Lancet (London, England)* 398.10294 (July 2021), pp. 27–40.

[15] Karina Kulangara et al. 'Clinical Utility of the Combined Positive Score for Programmed Death Ligand-1 Expression and the Approval of Pembrolizumab for Treatment of Gastric Cancer'. In: *Archives of Pathology & Laboratory Medicine* 143.3 (July 2018), pp. 330–337. (Visited on 24/09/2023).

[16] Dung T. Le et al. 'Mismatch Repair Deficiency Predicts Response of Solid Tumors to PD-1 Blockade'. In: *Science* 357.6349 (July 2017), pp. 409–413. (Visited on 24/09/2023).

[17] Yongfeng Wu et al. 'The Predictive Value of Tumor Mutation Burden on Efficacy of Immune Checkpoint Inhibitors in Cancers: A Systematic Review and Meta-Analysis'. In: *Frontiers in Oncology* 9 (2019). (Visited on 24/09/2023).

[18] Matthew Kyle Labriola et al. 'Characterization of Tumor Mutation Burden, PD-L1 and DNA Repair Genes to Assess Relationship to Immune Checkpoint Inhibitors Response in Metastatic Renal Cell Carcinoma'. In: *Journal for ImmunoTherapy of Cancer* 8.1 (Mar. 2020), e000319. (Visited on 24/09/2023).

[19] Jan Budczies et al. 'Quantifying Potential Confounders of Panel-Based Tumor Mutational Burden (TMB) Measurement'. In: *Lung Cancer* 142 (Apr. 2020), pp. 114–119. (Visited on 24/09/2023).

[20] Tae Hee Hong et al. 'Clinical Advantage of Targeted Sequencing for Unbiased Tumor Mutational Burden Estimation in Samples with Low Tumor Purity'. In: *Journal for ImmunoTherapy of Cancer* 8.2 (Oct. 2020), e001199. (Visited on 24/09/2023).

[21] Dorte Schou Nørøxe et al. 'Tumor Mutational Burden and Purity Adjustment before and after Treatment with Temozolomide in 27 Paired Samples of Glioblastoma: A Prospective Study'. In: *Molecular Oncology* 16.1 (2022), pp. 206–218. (Visited on 16/09/2023).

[22] Lorea Villanueva, Damiana Álvarez-Errico and Manel Esteller. 'The Contribution of Epigenetics to Cancer Immunotherapy'. In: *Trends in Immunology* 41.8 (Aug. 2020), pp. 676–691. (Visited on 23/09/2023).

[23] Michäel Duruisseaux et al. 'Epigenetic Prediction of Response to Anti-PD-1 Treatment in Non-Small-Cell Lung Cancer: A Multicentre, Retrospective Analysis'. In: *The Lancet. Respiratory Medicine* 6.10 (Oct. 2018), pp. 771–781.

[24] Ludwig Institute for Cancer Research. *Phase 1/2 Study of Anti-PD-L1 in Combination With Chemo(Radio)Therapy for Oesophageal Cancer*. Clinical Trial Registration NCT02735239. clinicaltrials.gov, June 2022. (Visited on 01/01/2023).

[25] *By Site | N. Ireland Cancer Registry*. https://www.qub.ac.uk/research-centres/nicr/CancerInformation/official-statistics/BySite/. (Visited on 10/07/2023).

[26] *Cancer Incidence in Wales, 2002-2019*. https://phw.nhs.wales/services-and-teams/welsh-cancer-intelligence-and-surveillance-unit-wcisu/cancer-incidence-in-wales-2002-2019/. (Visited on 10/07/2023).

[27] *Case-Mix Adjusted Percentage of Cancers Diagnosed at Stages 1 and 2 in England*. https://digital.nhs.uk/data-and-information/publications/statistical/case-mix-adjusted-percentage-of-cancers-diagnosed-at-stages-1-and-2-in-england. (Visited on 10/07/2023).

[28] Public Health Scotland. *Detect Cancer Early Staging Data - Year 10 and Impact of COVID-19*. https://publichealthscotland.scot/publications/detect-cancer-early-staging-data/detect-cancer-early-staging-data-year-10-and-impact-of-covid-19/. Statistical Report. (Visited on 10/07/2023).

[29] *Cancer Survival in England, Cancers Diagnosed 2016 to 2020, Followed up to 2021*. https://digital.nhs.uk/data-and-information/publications/statistical/cancer-survival-in-england/cancers-diagnosed-2016-to-2020-followed-up-to-2021. (Visited on 24/09/2023).

[30] Stuart Jon Spechler et al. 'History, Molecular Mechanisms, and Endoscopic Treatment of Barrett's Esophagus'. In: *Gastroenterology* 138.3 (Mar. 2010), pp. 854–869. (Visited on 25/09/2023).

[31] N. R. Barrett. 'Chronic Peptic Ulcer of the Oesophagus and 'Oesophagitis''. In: *The British Journal of Surgery* 38.150 (Oct. 1950), pp. 175–182.

[32] Lewis H. Bosher and Frederick H. Taylor. 'HETEROTOPIC GASTRIC MUCOSA IN THE ESOPHAGUS WITH ULCERATION AND STRICTURE FORMATION'. In: *Journal of Thoracic Surgery* 21.3 (Mar. 1951), pp. 306–312. (Visited on 25/09/2023).

[33] P. R. Allison and A. S. Johnstone. 'The Oesophagus Lined with Gastric Mucous Membrane'. In: *Thorax* 8.2 (June 1953), pp. 87–101. (Visited on 25/09/2023).

[34] Rodger C. Haggitt et al. 'Adenocarcinoma Complicating Columnar Epithelium-lined (Barrett's) Esophagus'. In: *American Journal of Clinical Pathology* 70.1 (July 1978), pp. 1–5. (Visited on 25/09/2023).

[35] Tarek Sawas et al. 'Identification of Prognostic Phenotypes of Esophageal Adenocarcinoma in 2 Independent Cohorts'. In: *Gastroenterology* 155.6 (Dec. 2018), 1720–1728.e4. (Visited on 26/09/2023).

[36] Karol Nowicki-Osuch et al. 'Molecular Phenotyping Reveals the Identity of Barrett's Esophagus and Its Malignant Transition'. In: *Science* 373.6556 (Aug. 2021), pp. 760–767. (Visited on 15/12/2023).

[37] Jianwen Que et al. 'Pathogenesis and Cells of Origin of Barrett's Esophagus'. In: *Gastroenterology* 157.2 (Aug. 2019), 349–364.e1. (Visited on 27/09/2023).

[38] Harit Kapoor et al. 'Animal Models of Barrett's Esophagus and Esophageal Adenocarcinoma–Past, Present, and Future'. In: *Clinical and Translational Science* 8.6 (Dec. 2015), pp. 841–847. (Visited on 26/09/2023).

[39] Rebecca C. Fitzgerald et al. 'British Society of Gastroenterology Guidelines on the Diagnosis and Management of Barrett's Oesophagus'. In: *Gut* 63.1 (Jan. 2014), pp. 7–42. (Visited on 27/09/2023).

[40] Michael Quante et al. 'Bile Acid and Inflammation Activate Gastric Cardia Stem Cells in a Mouse Model of Barrett-Like Metaplasia'. In: *Cancer Cell* 21.1 (Jan. 2012), pp. 36–51. (Visited on 27/09/2023).

[41] *Pubpeer Comment on Molecular Phenotyping Reveals the Identity of Barrett's Esophagus and Its Malignant Transition*. https://pubpeer.com/publications/43EE822A857DD67EC51DF9E8FEB35C. (Visited on 28/09/2023).

[42] Philippe Taniere et al. 'Cytokeratin Expression in Adenocarcinomas of the Esophagogastric Junction: A Comparative Study of Adenocarcinomas of the Distal Esophagus and of the Proximal Stomach'. In: *The American Journal of Surgical Pathology* 26.9 (Sept. 2002), p. 1213. (Visited on 28/09/2023).

[43] Xia Wang et al. 'Residual Embryonic Cells as Precursors of a Barrett's-like Metaplasia'. In: *Cell* 145.7 (June 2011), pp. 1023–1035. (Visited on 28/09/2023).

[44] Ming Jiang et al. 'Transitional Basal Cells at the Squamous–Columnar Junction Generate Barrett's Oesophagus'. In: *Nature* 550.7677 (Oct. 2017), pp. 529–533. (Visited on 29/09/2023).

[45] Richard I. Sherwood, Tzong-Yang Albert Chen and Douglas A. Melton. 'Transcriptional Dynamics of Endodermal Organ Formation'. In: *Developmental dynamics : an official publication of the American Association of Anatomists* 238.1 (Jan. 2009), pp. 29–42. (Visited on 29/09/2023).

[46] Jonathan N. Glickman et al. 'Multilayered Epithelium in Mucosal Biopsy Specimens From the Gastroesophageal Junction Region Is a Histologic Marker of Gastroesophageal Reflux Disease'. In: *The American Journal of Surgical Pathology* 33.6 (June 2009), p. 818. (Visited on 29/09/2023).

[47] Emil Goetsch. 'The Structure of the Mammalian Œsophagus'. In: *American Journal of Anatomy* 10.1 (1910), pp. 1–40. (Visited on 30/09/2023).

[48] Y. van Nieuwenhove, H. Destordeur and G. Willems. 'Spatial Distribution and Cell Kinetics of the Glands in the Human Esophageal Mucosa'. In: *European Journal of Morphology* 39.3 (July 2001), pp. 163–168.

[49] J G Azzopardi and T Menzies. 'Primary Œesophageal Adenocarcinoma. Confirmation of Its Existence by the Finding of Mucous Gland Tumours'. In: *British Journal of Surgery* 49.217 (Mar. 1962), pp. 497–506. (Visited on 30/09/2023).

[50] P Gillen et al. 'Experimental Columnar Metaplasia in the Canine Oesophagus'. In: *British Journal of Surgery* 75.2 (Feb. 1988), pp. 113–115. (Visited on 30/09/2023).

[51] H. Li et al. 'Mechanisms of Columnar Metaplasia and Squamous Regeneration in Experimental Barrett's Esophagus'. In: *Surgery* 115.2 (Feb. 1994), pp. 176–181.

[52] Rebecca A Coad et al. 'On the Histogenesis of Barrett's Oesophagus and Its Associated Squamous Islands: A Three-Dimensional Study of Their Morphological Relationship with Native Oesophageal Gland Ducts'. In: *The Journal of Pathology* 206.4 (2005), pp. 388–394. (Visited on 30/09/2023).

[53] S. J. Leedham et al. 'Individual Crypt Genetic Heterogeneity and the Origin of Metaplastic Glandular Epithelium in Human Barrett's Oesophagus'. In: *Gut* 57.8 (Aug. 2008), pp. 1041–1048. (Visited on 30/09/2023).

[54] Anna M. Nicholson et al. 'Barrett's Metaplasia Glands Are Clonal, Contain Multiple Stem Cells and Share a Common Squamous Progenitor'. In: *Gut* 61.10 (Oct. 2012), pp. 1380–1389. (Visited on 30/09/2023).

[55] Richard Peter Owen et al. 'Single Cell RNA-seq Reveals Profound Transcriptional Similarity between Barrett's Oesophagus and Oesophageal Submucosal Glands'. In: *Nature Communications* 9.1 (Oct. 2018), p. 4261. (Visited on 30/09/2023).

[56] W. Meyer, F. Vollmar and W. Bär. 'Barrett-Esophagus Following Total Gastrectomy. A Contribution to It's Pathogenesis'. In: *Endoscopy* 11.2 (May 1979), pp. 121–126.

[57] A. K. Sandvik and T. B. Halvorsen. 'Barrett's Esophagus after Total Gastrectomy'. In: *Journal of Clinical Gastroenterology* 10.5 (Oct. 1988), pp. 587–588.

[58] T. Tada et al. 'Adenocarcinoma Arising in Barrett's Esophagus after Total Gastrectomy'. In: *The American Journal of Gastroenterology* 85.11 (Nov. 1990), pp. 1503–1506.

[59] M. Konishi et al. 'Adenocarcinoma in Barrett's Esophagus Following Total Resection of the Gastric Remnant: A Case Report'. In: *Japanese Journal of Clinical Oncology* 22.4 (Aug. 1992), pp. 292–296.

[60] T. Nishimaki et al. 'Early Esophageal Adenocarcinoma Arising in a Short Segment of Barrett's Mucosa after Total Gastrectomy'. In: *The American Journal of Gastroenterology* 91.9 (Sept. 1996), pp. 1856–1857.

[61] Brenda C. Westhoff et al. 'Development of Barrett's Esophagus Six Months after Total Gastrectomy'. In: *The American Journal of Gastroenterology* 99.11 (Nov. 2004), pp. 2271–2277.

[62] Ulrich Peitz et al. 'Small-Bowel Metaplasia Arising in the Remnant Esophagus after Esophagojejunostomy–a [Corrected] Prospective Study in Patients with a History of Total Gastrectomy'. In: *The American Journal of Gastroenterology* 100.9 (Sept. 2005), pp. 2062–2070.

[63] Dong Hyun Sinn et al. 'Development of Barrett's Esophagus Soon after Total Gastrectomy'. In: *Gut and Liver* 2.1 (June 2008), pp. 51–53.

[64] Yutaka Shimada et al. 'Adenocarcinoma in Long-Segment Barrett's Esophagus 44 Years after Total Gastrectomy'. In: *Journal of Surgical Case Reports* 2013.12 (Dec. 2013), rjt100.

[65] Kailash Hemachandra and Douglas K. Rex. 'Development of Barrett's Esophagus after Total Gastrectomy'. In: *Gastrointestinal Endoscopy* 81.6 (2015), p. 1499.

[66] César Fernando Tróchez Mejía, Angelica Hernández Guerrero and Miguel Herrera Servin. 'Development of Barrett's Esophagus after a Total Gastrectomy'. In: *Revista Espanola De Enfermedades Digestivas* 112.12 (Dec. 2020), p. 954.

[67] B. A. E. Johns. 'Developmental Changes in the Oesophageal Epithelium in Man'. In: *Journal of Anatomy* 86.Pt 4 (Oct. 1952), pp. 431–442.4. (Visited on 30/09/2023).

[68] G. Gonzalez, Q. Huang and H. Mashimo. 'Characterization of Oncocytes in Deep Esophageal Glands'. In: *Diseases of the Esophagus* 29.6 (Sept. 2016), pp. 670–680. (Visited on 30/09/2023).

[69] J. H. Peters et al. 'Outcome of Adenocarcinoma Arising in Barrett's Esophagus in Endoscopically Surveyed and Nonsurveyed Patients'. In: *The Journal of Thoracic and Cardiovascular Surgery* 108.5 (Nov. 1994), 813–821, discussion 821–822.

[70] J. M. Streitz, C. W. Andrews and F. H. Ellis. 'Endoscopic Surveillance of Barrett's Esophagus. Does It Help?' In: *The Journal of Thoracic and Cardiovascular Surgery* 105.3 (Mar. 1993), 383–387, discussion 387–388.

[71] Gregory S. Cooper, Tzyung Doug Kou and Amitabh Chak. 'Receipt of Previous Diagnoses and Endoscopy and Outcome from Esophageal Adenocarcinoma: A Population-Based Study with Temporal Trends'. In: *The American Journal of Gastroenterology* 104.6 (June 2009), pp. 1356–1362.

[72] Pieter J. F. de Jonge et al. 'Risk of Malignant Progression in Patients with Barrett's Oesophagus: A Dutch Nationwide Cohort Study'. In: *Gut* 59.8 (Aug. 2010), pp. 1030–1036. (Visited on 25/09/2023).

[73] Frederik Hvid-Jensen et al. 'Incidence of Adenocarcinoma among Patients with Barrett's Esophagus'. In: *New England Journal of Medicine* 365.15 (Oct. 2011), pp. 1375–1383. (Visited on 25/09/2023).

[74] Oliver Old et al. 'Barrett's Oesophagus Surveillance versus Endoscopy at Need Study (BOSS): Protocol and Analysis Plan for a Multicentre Randomized Controlled Trial'. In: *Journal of Medical Screening* 22.3 (Sept. 2015), pp. 158–164.

[75] Sarah Killcoyne et al. 'Genomic Copy Number Predicts Esophageal Cancer Years before Transformation'. In: *Nature Medicine* 26.11 (Nov. 2020), pp. 1726–1732. (Visited on 15/12/2023).

[76] Douglas A. Corley et al. 'Surveillance and Survival in Barrett's Adenocarcinomas: A Population-Based Study'. In: *Gastroenterology* 122.3 (Mar. 2002), pp. 633–640.

[77] Gareth S. Dulai et al. 'Preoperative Prevalence of Barrett's Esophagus in Esophageal Adenocarcinoma: A Systematic Review'. In: *Gastroenterology* 122.1 (Jan. 2002), pp. 26–33.

[78] Sudarshan R Kadri et al. 'Acceptability and Accuracy of a Non-Endoscopic Screening Test for Barrett's Oesophagus in Primary Care: Cohort Study'. In: *The BMJ* 341 (Sept. 2010), p. c4372. (Visited on 02/10/2023).

[79] Caryn S. Ross-Innes et al. 'Evaluation of a Minimally Invasive Cell Sampling Device Coupled with Assessment of Trefoil Factor 3 Expression for Diagnosing Barrett's Esophagus: A Multi-Center Case–Control Study'. In: *PLoS Medicine* 12.1 (Jan. 2015), e1001780. (Visited on 02/10/2023).

[80] Rebecca C. Fitzgerald et al. 'Cytosponge-Trefoil Factor 3 versus Usual Care to Identify Barrett's Oesophagus in a Primary Care Setting: A Multicentre, Pragmatic, Randomised Controlled Trial'. In: *The Lancet* 396.10247 (Aug. 2020), pp. 333–344. (Visited on 02/10/2023).

[81] Anna L Paterson et al. 'Role of TFF3 as an Adjunct in the Diagnosis of Barrett's Esophagus Using a Minimally Invasive Esophageal Sampling Device – the Cytosponge™'. In: *Diagnostic cytopathology* 48.3 (Mar. 2020), pp. 253–264. (Visited on 02/10/2023).

[82] Evi S Lianidou et al. 'What's New on Circulating Tumor Cells? A Meeting Report'. In: *Breast Cancer Research : BCR* 12.4 (2010), p. 307. (Visited on 02/10/2023).

[83] Mina Nikanjam, Shumei Kato and Razelle Kurzrock. 'Liquid Biopsy: Current Technology and Clinical Applications'. In: *Journal of Hematology & Oncology* 15 (2022). (Visited on 02/10/2023).

[84] Ellen Heitzer, Lisa Auinger and Michael R. Speicher. 'Cell-Free DNA and Apoptosis: How Dead Cells Inform About the Living'. In: *Trends in Molecular Medicine* 26.5 (May 2020), pp. 519–528. (Visited on 21/12/2022).

[85] Maurice Stroun et al. 'Isolation and Characterization of DNA from the Plasma of Cancer Patients'. In: *European Journal of Cancer and Clinical Oncology* 23.6 (June 1987), pp. 707–712. (Visited on 02/10/2023).

[86] M. Stroun et al. 'Neoplastic Characteristics of the DNA Found in the Plasma of Cancer Patients'. In: *Oncology* 46.5 (1989), pp. 318–322.

[87] Homaira Nawroz et al. 'Microsatellite Alterations in Serum DNA of Head and Neck Cancer Patients'. In: *Nature Medicine* 2.9 (Sept. 1996), pp. 1035–1037. (Visited on 02/10/2023).

[88] M. Cisneros-Villanueva et al. 'Cell-Free DNA Analysis in Current Cancer Clinical Trials: A Review'. In: *British Journal of Cancer* 126.3 (Feb. 2022), pp. 391–400. (Visited on 22/12/2022).

[89] Arash Jamshidi et al. 'Evaluation of Cell-Free DNA Approaches for Multi-Cancer Early Detection'. In: *Cancer Cell* 40.12 (Dec. 2022), 1537–1549.e12. (Visited on 02/10/2023).

[90] Brian D. Nicholson et al. 'Multi-Cancer Early Detection Test in Symptomatic Patients Referred for Cancer Investigation in England and Wales (SYMPLIFY): A Large-Scale, Observational Cohort Study'. In: *The Lancet. Oncology* 24.7 (July 2023), pp. 733–743.

[91] Richard D. Neal et al. 'Cell-Free DNA-Based Multi-Cancer Early Detection Test in an Asymptomatic Screening Population (NHS-Galleri): Design of a Pragmatic, Prospective Randomised Controlled Trial'. In: *Cancers* 14.19 (Oct. 2022), p. 4818.

[92] T. Bestor et al. 'Cloning and Sequencing of a cDNA Encoding DNA Methyltransferase of Mouse Cells. The Carboxyl-Terminal Domain of the Mammalian Enzymes Is Related to Bacterial Restriction Methyltransferases'. In: *Journal of Molecular Biology* 203.4 (Oct. 1988), pp. 971–983.

[93] J. D. Smith and Roy Markham. 'The Enzymic Breakdown of Deoxyribosenucleic Acids'. In: *Biochimica et Biophysica Acta* 8 (Jan. 1952), pp. 350–351. (Visited on 10/10/2023).

[94] J. D. Smith and Roy Markham. 'Polynucleotides from Deoxyribonucleic Acids'. In: *Nature* 170.4316 (July 1952), pp. 120–121. (Visited on 10/10/2023).

[95] Jikui Song et al. 'Structure-Based Mechanistic Insights into DNMT1-Mediated Maintenance DNA Methylation'. In: *Science* 335.6069 (Feb. 2012), pp. 709–712. (Visited on 04/10/2023).

[96] Jikui Song et al. 'Structure of DNMT1-DNA Complex Reveals a Role for Autoinhibition in Maintenance DNA Methylation'. In: *Science* 331.6020 (Feb. 2011), pp. 1036–1040. (Visited on 15/12/2023).

[97] Kohei Takeshita et al. 'Structural Insight into Maintenance Methylation by Mouse DNA Methyltransferase 1 (Dnmt1)'. In: *Proceedings of the National Academy of Sciences* 108.22 (May 2011), pp. 9055–9059. (Visited on 04/10/2023).

[98] Monica Mancini et al. 'The Multi-Functionality of UHRF1: Epigenome Maintenance and Preservation of Genome Integrity'. In: *Nucleic Acids Research* 49.11 (June 2021), pp. 6053–6068. (Visited on 05/10/2023).

[99] Magnolia Bostick et al. 'UHRF1 Plays a Role in Maintaining DNA Methylation in Mammalian Cells'. In: *Science* 317.5845 (Sept. 2007), pp. 1760–1764. (Visited on 04/10/2023).

[100] Jafar Sharif et al. 'The SRA Protein Np95 Mediates Epigenetic Inheritance by Recruiting Dnmt1 to Methylated DNA'. In: *Nature* 450.7171 (Dec. 2007), pp. 908–912. (Visited on 04/10/2023).

[101] Scott B. Rothbart et al. 'Multivalent Histone Engagement by the Linked Tandem Tudor and PHD Domains of UHRF1 Is Required for the Epigenetic Inheritance of DNA Methylation'. In: *Genes & Development* 27.11 (June 2013), pp. 1288–1298. (Visited on 06/10/2023).

[102] Satoshi Ishiyama et al. 'Structure of the Dnmt1 Reader Module Complexed with a Unique Two-Mono-Ubiquitin Mark on Histone H3 Reveals the Basis for DNA Methylation Maintenance'. In: *Molecular Cell* 68.2 (Oct. 2017), 350–360.e7. (Visited on 05/10/2023).

[103] Weihua Qin et al. 'DNA Methylation Requires a DNMT1 Ubiquitin Interacting Motif (UIM) and Histone Ubiquitination'. In: *Cell Research* 25.8 (Aug. 2015), pp. 911–929. (Visited on 05/10/2023).

[104] Atsuya Nishiyama et al. 'Uhrf1-Dependent H3K23 Ubiquitylation Couples Maintenance DNA Methylation and Replication'. In: *Nature* 502.7470 (Oct. 2013), pp. 249–253. (Visited on 05/10/2023).

[105] Giedrius Vilkaitis et al. 'Processive Methylation of Hemimethylated CpG Sites by Mouse Dnmt1 DNA Methyltransferase*'. In: *Journal of Biological Chemistry* 280.1 (Jan. 2005), pp. 64–72. (Visited on 05/10/2023).

[106] Ronald Garingalao Garvilles et al. 'Dual Functions of the RFTS Domain of Dnmt1 in Replication-Coupled DNA Methylation and in Protection of the Genome from Aberrant Methylation'. In: *PLOS ONE* 10.9 (Sept. 2015), e0137509. (Visited on 04/10/2023).

[107] Yan Wang, Huijie Liu and Zhongsheng Sun. 'Lamarck Rises from His Grave: Parental Environment-Induced Epigenetic Inheritance in Model Organisms and Humans'. In: *Biological Reviews* 92.4 (2017), pp. 2084–2111. (Visited on 06/10/2023).

[108] Masaki Okano, Shaoping Xie and En Li. 'Cloning and Characterization of a Family of Novel Mammalian DNA (Cytosine-5) Methyltransferases'. In: *Nature Genetics* 19.3 (July 1998), pp. 219–220. (Visited on 05/10/2023).

[109] Zhi-Min Zhang et al. 'Structural Basis for DNMT3A-mediated de Novo DNA Methylation'. In: *Nature* 554.7692 (Feb. 2018), pp. 387–391. (Visited on 05/10/2023).

[110] Déborah Bourc'his et al. 'Dnmt3L and the Establishment of Maternal Genomic Imprints'. In: *Science* 294.5551 (Dec. 2001), pp. 2536–2539. (Visited on 06/10/2023).

[111] Christopher E. Duymich et al. 'DNMT3B Isoforms without Catalytic Activity Stimulate Gene Body Methylation as Accessory Proteins in Somatic Cells'. In: *Nature Communications* 7.1 (Apr. 2016), p. 11453. (Visited on 06/10/2023).

[112] Daniel J. Weisenberger et al. 'Role of the DNA Methyltransferase Variant DNMT3b3 in DNA Methylation1'. In: *Molecular Cancer Research* 2.1 (Jan. 2004), pp. 62–72. (Visited on 06/10/2023).

[113] Xue Guo et al. 'Structural Insight into Autoinhibition and Histone H3-induced Activation of DNMT3A'. In: *Nature* 517.7536 (Jan. 2015), pp. 640–644. (Visited on 06/10/2023).

[114] Liubin Yang, Rachel Rau and Margaret A. Goodell. 'DNMT3A in Haematological Malignancies'. In: *Nature Reviews Cancer* 15.3 (Mar. 2015), pp. 152–165. (Visited on 07/10/2023).

[115] R. Scott Hansen et al. 'The DNMT3B DNA Methyltransferase Gene Is Mutated in the ICF Immunodeficiency Syndrome'. In: *Proceedings of the National Academy of Sciences* 96.25 (Dec. 1999), pp. 14412–14417. (Visited on 07/10/2023).

[116] Bethany L. Wienholz et al. 'DNMT3L Modulates Significant and Distinct Flanking Sequence Preference for DNA Methylation by DNMT3A and DNMT3B In Vivo'. In: *PLOS Genetics* 6.9 (Sept. 2010), e1001106. (Visited on 07/10/2023).

[117] Linfeng Gao et al. 'Comprehensive Structure-Function Characterization of DNMT3B and DNMT3A Reveals Distinctive de Novo DNA Methylation Mechanisms'. In: *Nature Communications* 11.1 (July 2020), p. 3355. (Visited on 06/10/2023).

[118] Allison B. Norvil et al. 'Dnmt3b Methylates DNA by a Noncooperative Mechanism, and Its Activity Is Unaffected by Manipulations at the Predicted Dimer Interface'. In: *Biochemistry* 57.29 (July 2018), pp. 4312–4324. (Visited on 07/10/2023).

[119] Chien-Chu Lin et al. 'Structural Insights into CpG-specific DNA Methylation by Human DNA Methyltransferase 3B'. In: *Nucleic Acids Research* 48.7 (Apr. 2020), pp. 3949–3961. (Visited on 07/10/2023).

[120] Humaira Gowher and Albert Jeltsch. 'Molecular Enzymology of the Catalytic Domains of the Dnmt3a and Dnmt3b DNA Methyltransferases*'. In: *Journal of Biological Chemistry* 277.23 (June 2002), pp. 20409–20414. (Visited on 07/10/2023).

[121] Renata Z. Jurkowska et al. 'Formation of Nucleoprotein Filaments by Mammalian DNA Methyltransferase Dnmt3a in Complex with Regulator Dnmt3L'. In: *Nucleic Acids Research* 36.21 (Dec. 2008), pp. 6656–6663. (Visited on 07/10/2023).

[122] Renata Z. Jurkowska et al. 'Oligomerization and Binding of the Dnmt3a DNA Methyltransferase to Parallel DNA Molecules: HETEROCHROMATIC LOCALIZATION AND ROLE OF Dnmt3L*'. In: *Journal of Biological Chemistry* 286.27 (July 2011), pp. 24200–24207. (Visited on 07/10/2023).

[123] Arumugam Rajavelu et al. 'Function and Disruption of DNA Methyltransferase 3a Cooperative DNA Binding and Nucleoprotein Filament Formation'. In: *Nucleic Acids Research* 40.2 (Jan. 2012), pp. 569–580. (Visited on 07/10/2023).

[124] Susan C. Wu and Yi Zhang. 'Active DNA Demethylation: Many Roads Lead to Rome'. In: *Nature Reviews Molecular Cell Biology* 11.9 (Sept. 2010), pp. 607–620. (Visited on 08/10/2023).

[125] Shinsuke Ito et al. 'Tet Proteins Can Convert 5-Methylcytosine to 5-Formylcytosine and 5-Carboxylcytosine'. In: *Science* 333.6047 (Sept. 2011), pp. 1300–1303. (Visited on 08/10/2023).

[126] Yu-Fei He et al. 'Tet-Mediated Formation of 5-Carboxylcytosine and Its Excision by TDG in Mammalian DNA'. In: *Science* 333.6047 (Sept. 2011), pp. 1303–1307. (Visited on 08/10/2023).

[127] Atanu Maiti and Alexander C. Drohat. 'Thymine DNA Glycosylase Can Rapidly Excise 5-Formylcytosine and 5-Carboxylcytosine: POTENTIAL IMPLICATIONS FOR ACTIVE DEMETHYLATION OF CpG SITES*'. In: *Journal of Biological Chemistry* 286.41 (Oct. 2011), pp. 35334–35338. (Visited on 08/10/2023).

[128] Hideharu Hashimoto et al. 'Recognition and Potential Mechanisms for Replication and Erasure of Cytosine Hydroxymethylation'. In: *Nucleic Acids Research* 40.11 (June 2012), pp. 4841–4849. (Visited on 09/10/2023).

[129] Lulu Hu et al. 'Structural Insight into Substrate Preference for TET-mediated Oxidation'. In: *Nature* 527.7576 (Nov. 2015), pp. 118–122. (Visited on 08/10/2023).

[130] Yang Zeng and Taiping Chen. 'DNA Methylation Reprogramming during Mammalian Development'. In: *Genes* 10.4 (Apr. 2019), p. 257. (Visited on 08/10/2023).

[131] Hannah K. Long, Neil P. Blackledge and Robert J. Klose. 'ZF-CxxC Domain-Containing Proteins, CpG Islands and the Chromatin Connection'. In: *Biochemical Society Transactions* 41.3 (May 2013), pp. 727–740. (Visited on 07/10/2023).

[132] Chao Xu et al. 'DNA Sequence Recognition of Human CXXC Domains and Their Structural Determinants'. In: *Structure* 26.1 (Jan. 2018), 85–95.e3. (Visited on 08/10/2023).

[133] Myunggon Ko et al. 'Modulation of TET2 Expression and 5-Methylcytosine Oxidation by the CXXC Domain Protein IDAX'. In: *Nature* 497.7447 (May 2013), pp. 122–126. (Visited on 08/10/2023).

[134] Lulu Hu et al. 'Crystal Structure of TET2-DNA Complex: Insight into TET-Mediated 5mC Oxidation'. In: *Cell* 155.7 (Dec. 2013), pp. 1545–1555. (Visited on 08/10/2023).

[135] Alain R. Weber et al. 'Biochemical Reconstitution of TET1–TDG–BER-dependent Active DNA Demethylation Reveals a Highly Coordinated Mechanism'. In: *Nature Communications* 7.1 (Mar. 2016), p. 10806. (Visited on 08/10/2023).

[136] Xiaoping Liu et al. 'UHRF2 Commissions the Completion of DNA Demethylation through Allosteric Activation by 5hmC and K33-linked Ubiquitination of XRCC1'. In: *Molecular Cell* 81.14 (July 2021), 2960–2974.e7. (Visited on 08/10/2023).

[137] Cornelia G. Spruijt et al. 'Dynamic Readers for 5-(Hydroxy)Methylcytosine and Its Oxidized Derivatives'. In: *Cell* 152.5 (Feb. 2013), pp. 1146–1159. (Visited on 08/10/2023).

[138] Ting Zhou et al. 'Structural Basis for Hydroxymethylcytosine Recognition by the SRA Domain of UHRF2'. In: *Molecular Cell* 54.5 (June 2014), pp. 879–886. (Visited on 08/10/2023).

[139] Robert M Vaughan et al. 'Comparative Biochemical Analysis of UHRF Proteins Reveals Molecular Mechanisms That Uncouple UHRF2 from DNA Methylation Maintenance'. In: *Nucleic Acids Research* 46.9 (May 2018), pp. 4405–4416. (Visited on 08/10/2023).

[140] Shane M. Ginnard et al. 'Molecular Investigation of the Tandem Tudor Domain and Plant Homeodomain Histone Binding Domains of the Epigenetic Regulator UHRF2'. In: *Proteins* 90.3 (Mar. 2022), pp. 835–847.

[141] Toshinobu Nakamura et al. 'PGC7 Binds Histone H3K9me2 to Protect against Conversion of 5mC to 5hmC in Early Embryos'. In: *Nature* 486.7403 (June 2012), pp. 415–419. (Visited on 09/10/2023).

[142] Yingfeng Li et al. 'Stella Safeguards the Oocyte Methylome by Preventing de Novo Methylation Mediated by DNMT1'. In: *Nature* 564.7734 (Dec. 2018), pp. 136–140. (Visited on 09/10/2023).

[143] Rollin D. Hotchkiss. 'THE QUANTITATIVE SEPARATION OF PURINES, PYRIMIDINES, AND NUCLEOSIDES BY PAPER CHROMATOGRAPHY'. In: *Journal of Biological Chemistry* 175.1 (Aug. 1948), pp. 315–332. (Visited on 09/10/2023).

[144] B. F. Vanyushin, S. G. Tkacheva and A. N. Belozersky. 'Rare Bases in Animal DNA'. In: *Nature* 225.5236 (Mar. 1970), pp. 948–949. (Visited on 09/10/2023).

[145] Judith Singer, Joan Roberts-Ems and Arthur D. Riggs. 'Methylation of Mouse Liver DNA Studied by Means of the Restriction Enzymes Msp I and Hpa II'. In: *Science* 203.4384 (Mar. 1979), pp. 1019–1021. (Visited on 05/10/2023).

[146] C. Waalwijk and R.A. Flavell. 'MspI, an Isoschizomer of HpaII Which Cleaves Both Unmethylated and Methylated HpaII Sites'. In: *Nucleic Acids Research* 5.9 (Sept. 1978), pp. 3231–3236. (Visited on 05/10/2023).

[147] M Frommer et al. 'A Genomic Sequencing Protocol That Yields a Positive Display of 5-Methylcytosine Residues in Individual DNA Strands.' In: *Proceedings of the National Academy of Sciences* 89.5 (Mar. 1992), pp. 1827–1831. (Visited on 09/10/2023).

[148] Yibin Liu et al. 'Bisulfite-Free Direct Detection of 5-Methylcytosine and 5-Hydroxymethylcytosine at Base Resolution'. In: *Nature Biotechnology* 37.4 (Apr. 2019), pp. 424–429.

[149] Romualdas Vaisvila et al. 'Enzymatic Methyl Sequencing Detects DNA Methylation at Single-Base Resolution from Picograms of DNA'. In: *Genome Research* 31.7 (July 2021), pp. 1280–1289. (Visited on 09/10/2023).

[150] Tong Wang et al. 'Direct Enzymatic Sequencing of 5-Methylcytosine at Single-Base Resolution'. In: *Nature Chemical Biology* 19.8 (Aug. 2023), pp. 1004–1012. (Visited on 09/10/2023).

[151] Michael Weber et al. 'Chromosome-Wide and Promoter-Specific Analyses Identify Sites of Differential DNA Methylation in Normal and Transformed Human Cells'. In: *Nature Genetics* 37.8 (Aug. 2005), pp. 853–862. (Visited on 09/10/2023).

[152] Yang Liu et al. 'DNA Methylation-Calling Tools for Oxford Nanopore Sequencing: A Survey and Human Epigenome-Wide Evaluation'. In: *Genome Biology* 22.1 (Oct. 2021), p. 295. (Visited on 09/10/2023).

[153] Benjamin A. Flusberg et al. 'Direct Detection of DNA Methylation during Single-Molecule, Real-Time Sequencing'. In: *Nature Methods* 7.6 (June 2010), pp. 461–465. (Visited on 09/10/2023).

[154] S. K. Myöhänen, S. B. Baylin and J. G. Herman. 'Hypermethylation Can Selectively Silence Individual p16ink4A Alleles in Neoplasia'. In: *Cancer Research* 58.4 (Feb. 1998), pp. 591–593.

[155] J G Herman et al. 'Silencing of the VHL Tumor-Suppressor Gene by DNA Methylation in Renal Carcinoma.' In: *Proceedings of the National Academy of Sciences* 91.21 (Oct. 1994), pp. 9700–9704. (Visited on 11/10/2023).

[156] James G. Herman et al. 'Incidence and Functional Consequences of hMLH1 Promoter Hypermethylation in Colorectal Carcinoma'. In: *Proceedings of the National Academy of Sciences* 95.12 (June 1998), pp. 6870–6875. (Visited on 11/10/2023).

[157] Manel Esteller et al. 'Promoter Hypermethylation and BRCA1 Inactivation in Sporadic Breast and Ovarian Tumors'. In: *JNCI: Journal of the National Cancer Institute* 92.7 (Apr. 2000), pp. 564–569. (Visited on 11/10/2023).

[158] R M Liskay and R J Evans. 'Inactive X Chromosome DNA Does Not Function in DNA-mediated Cell Transformation for the Hypoxanthine Phosphoribosyltransferase Gene.' In: *Proceedings of the National Academy of Sciences* 77.8 (Aug. 1980), pp. 4895–4898. (Visited on 10/10/2023).

[159] A Fradin, J L Manley and C L Prives. 'Methylation of Simian Virus 40 Hpa II Site Affects Late, but Not Early, Viral Gene Expression.' In: *Proceedings of the National Academy of Sciences* 79.17 (Sept. 1982), pp. 5142–5146. (Visited on 10/10/2023).

[160] F. Watt and P. L. Molloy. 'Cytosine Methylation Prevents Binding to DNA of a HeLa Cell Transcription Factor Required for Optimal Expression of the Adenovirus Major Late Promoter.' In: *Genes & Development* 2.9 (Sept. 1988), pp. 1136–1143. (Visited on 10/10/2023).

[161] Richard R. Meehan et al. 'Identification of a Mammalian Protein That Binds Specifically to DNA Containing Methylated CpGs'. In: *Cell* 58.3 (Aug. 1989), pp. 499–507. (Visited on 10/10/2023).

[162] Joan Boyes and Adrian Bird. 'DNA Methylation Inhibits Transcription Indirectly via a Methyl-CpG Binding Protein'. In: *Cell* 64.6 (Mar. 1991), pp. 1123–1134. (Visited on 10/10/2023).

[163] Richard Meehan, Joe D. Lewis and Adrian P. Bird. 'Characterization of MeCP2, a Vertebrate DNA Binding Protein with Affinity for Methylated DNA'. In: *Nucleic Acids Research* 20.19 (Oct. 1992), pp. 5085–5092. (Visited on 10/10/2023).

[164] P H Yen et al. 'Differential Methylation of Hypoxanthine Phosphoribosyltransferase Genes on Active and Inactive Human X Chromosomes.' In: *Proceedings of the National Academy of Sciences* 81.6 (Mar. 1984), pp. 1759–1763. (Visited on 10/10/2023).

[165] Adrian Bird et al. 'A Fraction of the Mouse Genome That Is Derived from Islands of Nonmethylated, CpG-rich DNA'. In: *Cell* 40.1 (Jan. 1985), pp. 91–99. (Visited on 10/10/2023).

[166] Adrian P. Bird. 'DNA Methylation and the Frequency of CpG in Animal DNA'. In: *Nucleic Acids Research* 8.7 (Apr. 1980), pp. 1499–1504. (Visited on 10/10/2023).

[167] D. Barker, M. Schafer and R. White. 'Restriction Sites Containing CpG Show a Higher Frequency of Polymorphism in Human DNA'. In: *Cell* 36.1 (Jan. 1984), pp. 131–138.

[168] Adrian P. Bird. 'CpG-rich Islands and the Function of DNA Methylation'. In: *Nature* 321.6067 (May 1986), pp. 209–213. (Visited on 10/10/2023).

[169] Amy T. Hark et al. 'CTCF Mediates Methylation-Sensitive Enhancer-Blocking Activity at the H19/Igf2 Locus'. In: *Nature* 405.6785 (May 2000), pp. 486–489. (Visited on 11/10/2023).

[170] Yan Li et al. 'The Structural Basis for Cohesin–CTCF-anchored Loops'. In: *Nature* 578.7795 (Feb. 2020), pp. 472–476. (Visited on 11/10/2023).

[171] Robert A. Drewell et al. 'Deletion of a Silencer Element Disrupts H19 Imprinting Independently of a DNA Methylation Epigenetic Switch'. In: *Development* 127.16 (Aug. 2000), pp. 3419–3428. (Visited on 11/10/2023).

[172] Karin Buiting, Charles Williams and Bernhard Horsthemke. 'Angelman Syndrome — Insights into a Rare Neurogenetic Disorder'. In: *Nature Reviews Neurology* 12.10 (Oct. 2016), pp. 584–593. (Visited on 11/10/2023).

[173] Linyan Meng et al. 'Truncation of Ube3a-ATS Unsilences Paternal Ube3a and Ameliorates Behavioral Defects in the Angelman Syndrome Mouse Model'. In: *PLOS Genetics* 9.12 (Dec. 2013), e1004039. (Visited on 11/10/2023).

[174] Linyan Meng et al. 'Towards a Therapy for Angelman Syndrome by Targeting a Long Non-Coding RNA'. In: *Nature* 518.7539 (Feb. 2015), pp. 409–412. (Visited on 11/10/2023).

[175] Xiao Li and Xiang-Dong Fu. 'Chromatin-Associated RNAs as Facilitators of Functional Genomic Interactions'. In: *Nature Reviews Genetics* 20.9 (Sept. 2019), pp. 503–519. (Visited on 11/10/2023).

[176] Peter A. Jones. 'The DNA Methylation Paradox'. In: *Trends in Genetics* 15.1 (Jan. 1999), pp. 34–37. (Visited on 12/10/2023).

[177] Tuncay Baubec et al. 'Genomic Profiling of DNA Methyltransferases Reveals a Role for DNMT3B in Genic Methylation'. In: *Nature* 520.7546 (Apr. 2015), pp. 243–247. (Visited on 12/10/2023).

[178] Francesco Neri et al. 'Intragenic DNA Methylation Prevents Spurious Transcription Initiation'. In: *Nature* 543.7643 (Mar. 2017), pp. 72–77. (Visited on 09/10/2023).

[179] Eric J. Wagner and Phillip B. Carpenter. 'Understanding the Language of Lys36 Methylation at Histone H3'. In: *Nature reviews. Molecular cell biology* 13.2 (Jan. 2012), pp. 115–126. (Visited on 12/10/2023).

[180] Arunkumar Dhayalan et al. 'The Dnmt3a PWWP Domain Reads Histone 3 Lysine 36 Trimethylation and Guides DNA Methylation*'. In: *Journal of Biological Chemistry* 285.34 (Aug. 2010), pp. 26114–26120. (Visited on 12/10/2023).

[181] L. Arce, N. N. Yokoyama and M. L. Waterman. 'Diversity of LEF/TCF Action in Development and Disease'. In: *Oncogene* 25.57 (Dec. 2006), pp. 7492–7504. (Visited on 12/10/2023).

[182] Tony W.-H. Li et al. 'Wnt Activation and Alternative Promoter Repression of LEF1 in Colon Cancer'. In: *Molecular and Cellular Biology* 26.14 (July 2006), pp. 5284–5299. (Visited on 12/10/2023).

[183] Marjolijn J. L. Ligtenberg et al. 'Heritable Somatic Methylation and Inactivation of MSH2 in Families with Lynch Syndrome Due to Deletion of the 3′ Exons of TACSTD1'. In: *Nature Genetics* 41.1 (Jan. 2009), pp. 112–117. (Visited on 11/10/2023).

[184] V. M. Barbour et al. 'Alpha-Thalassemia Resulting from a Negative Chromosomal Position Effect'. In: *Blood* 96.3 (Aug. 2000), pp. 800–807.

[185] Hannah K Long et al. 'Epigenetic Conservation at Gene Regulatory Elements Revealed by Non-Methylated DNA Profiling in Seven Vertebrates'. In: *eLife* 2 (Feb. 2013). Ed. by Anne Ferguson-Smith, e00348. (Visited on 12/10/2023).

[186] Bradley E. Bernstein et al. 'A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells'. In: *Cell* 125.2 (Apr. 2006), pp. 315–326. (Visited on 12/10/2023).

[187] Wei Xie et al. 'Epigenomic Analysis of Multilineage Differentiation of Human Embryonic Stem Cells'. In: *Cell* 153.5 (May 2013), pp. 1134–1148. (Visited on 12/10/2023).

[188] Mira Jeong et al. 'Large Conserved Domains of Low DNA Methylation Maintained by Dnmt3a'. In: *Nature Genetics* 46.1 (Jan. 2014), pp. 17–23. (Visited on 12/10/2023).

[189] Alexander Meissner et al. 'Genome-Scale DNA Methylation Maps of Pluripotent and Differentiated Cells'. In: *Nature* 454.7205 (Aug. 2008), pp. 766–770. (Visited on 12/10/2023).

[190] Einav Nili Gal-Yam et al. 'Frequent Switching of Polycomb Repressive Marks and DNA Hypermethylation in the PC3 Prostate Cancer Cell Line'. In: *Proceedings of the National Academy of Sciences of the United States of America* 105.35 (Sept. 2008), pp. 12979–12984.

[191] Andrew E. Teschendorff et al. 'Age-Dependent DNA Methylation of Genes That Are Suppressed in Stem Cells Is a Hallmark of Cancer'. In: *Genome Research* 20.4 (Apr. 2010), pp. 440–446. (Visited on 12/10/2023).

[192] Ashraf Dallol et al. 'Methylation of the Polycomb Group Target Genes Is a Possible Biomarker for Favorable Prognosis in Colorectal Cancer'. In: *Cancer Epidemiology, Biomarkers & Prevention* 21.11 (Nov. 2012), pp. 2069–2075. (Visited on 12/10/2023).

[193] Minoru Toyota et al. 'CpG Island Methylator Phenotype in Colorectal Cancer'. In: *Proceedings of the National Academy of Sciences of the United States of America* 96.15 (July 1999), pp. 8681–8686. (Visited on 12/10/2023).

[194] Orlando J. Miller et al. '5-Methylcytosine Localised in Mammalian Constitutive Heterochromatin'. In: *Nature* 251.5476 (Oct. 1974), pp. 636–637. (Visited on 12/10/2023).

[195] J. D. Lewis et al. 'Purification, Sequence, and Cellular Localization of a Novel Chromosomal Protein That Binds to Methylated DNA'. In: *Cell* 69.6 (June 1992), pp. 905–914.

[196] Judith Singer-Sam and Arthur D. Riggs. 'X Chromosome Inactivation and DNA Methylation'. In: *DNA Methylation: Molecular Biology and Biological Significance*. Ed. by Jean-Pierre Jost and Hans-Peter Saluz. EXS. Basel: Birkhäuser, 1993, pp. 358–384. (Visited on 12/10/2023).

[197] Andrea Rottach et al. 'The Multi-Domain Protein Np95 Connects DNA Methylation and Histone Modification'. In: *Nucleic Acids Research* 38.6 (Apr. 2010), pp. 1796–1804. (Visited on 12/10/2023).

[198] Asaf Hellman and Andrew Chess. 'Gene Body-Specific Methylation on the Active X Chromosome'. In: *Science* (Feb. 2007). (Visited on 11/01/2022).

[199] Ryan Lister et al. 'Human DNA Methylomes at Base Resolution Show Widespread Epigenomic Differences'. In: *Nature* 462.7271 (Nov. 2009), pp. 315–322. (Visited on 12/10/2023).

[200] R. L. P. Adams. 'The Relationship between Synthesis and Methylation of DNA in Mouse Fibroblasts'. In: *Biochimica et Biophysica Acta (BBA) - Nucleic Acids and Protein Synthesis* 254.2 (Dec. 1971), pp. 205–212. (Visited on 13/10/2023).

[201] Andrew P. Feinberg and Bert Vogelstein. 'Hypomethylation Distinguishes Genes of Some Human Cancers from Their Normal Counterparts'. In: *Nature* 301.5895 (Jan. 1983), pp. 89–92. (Visited on 13/10/2023).

[202] M A Gama-Sosa et al. 'The 5-Methylcytosine Content of DNA from Human Tumors.' In: *Nucleic Acids Research* 11.19 (Oct. 1983), pp. 6883–6894. (Visited on 13/10/2023).

[203] Bo Wen et al. 'Large Histone H3 Lysine 9 Dimethylated Chromatin Blocks Distinguish Differentiated from Embryonic Stem Cells'. In: *Nature Genetics* 41.2 (Feb. 2009), pp. 246–250. (Visited on 13/10/2023).

[204] Kasper Daniel Hansen et al. 'Increased Methylation Variation in Epigenetic Domains across Cancer Types'. In: *Nature Genetics* 43.8 (Aug. 2011), pp. 768–775. (Visited on 11/01/2022).

[205] Winston Timp et al. 'Large Hypomethylated Blocks as a Universal Defining Epigenetic Alteration in Human Solid Tumors'. In: *Genome Medicine* 6.8 (Aug. 2014), p. 61. (Visited on 25/07/2023).

[206] Amy R. Vandiver et al. 'Age and Sun Exposure-Related Widespread Genomic Blocks of Hypomethylation in Nonmalignant Skin'. In: *Genome Biology* 16.1 (Apr. 2015), p. 80. (Visited on 13/10/2023).

[207] Benjamin P. Berman et al. 'Regions of Focal DNA Hypermethylation and Long-Range Hypomethylation in Colorectal Cancer Coincide with Nuclear Lamina–Associated Domains'. In: *Nature Genetics* 44.1 (Jan. 2012), pp. 40–46. (Visited on 25/07/2023).

[208] Hazel A. Cruickshanks et al. 'Senescent Cells Harbour Features of the Cancer Epigenome'. In: *Nature Cell Biology* 15.12 (Dec. 2013), pp. 1495–1506. (Visited on 13/10/2023).

[209] Wanding Zhou et al. 'DNA Methylation Loss in Late-Replicating Domains Is Linked to Mitotic Cell Division'. In: *Nature Genetics* 50.4 (Apr. 2018), pp. 591–602. (Visited on 25/07/2023).

[210] Atsuya Nishiyama et al. 'Two Distinct Modes of DNMT1 Recruitment Ensure Stable Maintenance DNA Methylation'. In: *Nature Communications* 11.1 (Mar. 2020), p. 1222. (Visited on 04/10/2023).

[211] Thomas J. Hudson (Chairperson) et al. 'International Network of Cancer Genome Projects'. In: *Nature* 464.7291 (Apr. 2010), pp. 993–998. (Visited on 15/10/2023).

[212] John N. Weinstein et al. 'The Cancer Genome Atlas Pan-Cancer Analysis Project'. In: *Nature Genetics* 45.10 (Oct. 2013), pp. 1113–1120. (Visited on 15/10/2023).

[213] Brian J. Reid et al. 'Barrett's Esophagus: Correlation between Flow Cytometry and Histology in Detection of Patients at Risk for Adenocarcinoma'. In: *Gastroenterology* 93.1 (July 1987), pp. 1–11. (Visited on 22/09/2023).

[214] Caryn S. Ross-Innes et al. 'Whole-Genome Sequencing Provides New Insights into the Clonal Architecture of Barrett's Esophagus and Esophageal Adenocarcinoma'. In: *Nature Genetics* 47.9 (Sept. 2015), pp. 1038–1046. (Visited on 14/10/2023).

[215] Emily G. Barr Fritcher et al. 'A Comparison of Conventional Cytology, DNA Ploidy Analysis, and Fluorescence in Situ Hybridization for the Detection of Dysplasia and Adenocarcinoma in Patients with Barrett's Esophagus'. In: *Human Pathology* 39.8 (Aug. 2008), pp. 1128–1135.

[216] Won-Tak Choi et al. 'Diagnosis and Risk Stratification of Barrett's Dysplasia by Flow Cytometric DNA Analysis of Paraffin-Embedded Tissue'. In: *Gut* 67.7 (July 2018), pp. 1229–1238.

[217] Matthew D. Stachler et al. 'Paired Exome Analysis of Barrett's Esophagus and Adenocarcinoma'. In: *Nature Genetics* 47.9 (Sept. 2015), pp. 1047–1055. (Visited on 13/10/2023).

[218] Katia Nones et al. 'Genomic Catastrophes Frequently Arise in Esophageal Adenocarcinoma and Drive Tumorigenesis'. In: *Nature Communications* 5.1 (Oct. 2014), p. 5224. (Visited on 13/10/2023).

[219] Jens Luebeck et al. 'Extrachromosomal DNA in the Cancerous Transformation of Barrett's Oesophagus'. In: *Nature* 616.7958 (Apr. 2023), pp. 798–805. (Visited on 14/10/2023).

[220] Joshua T. Lange et al. 'The Evolutionary Dynamics of Extrachromosomal DNA in Human Cancers'. In: *Nature Genetics* 54.10 (Oct. 2022), pp. 1527–1533. (Visited on 14/10/2023).

[221] Austin M. Dulak et al. 'Gastrointestinal Adenocarcinomas of the Esophagus, Stomach, and Colon Exhibit Distinct Patterns of Genome Instability and Oncogenesis'. In: *Cancer Research* 72.17 (Aug. 2012), pp. 4383–4393. (Visited on 14/10/2023).

[222] Jihun Kim et al. 'Integrated Genomic Characterization of Oesophageal Carcinoma'. In: *Nature* 541.7636 (Jan. 2017), pp. 169–175. (Visited on 16/08/2023).

[223] Alexander M. Frankell et al. 'The Landscape of Selection in 551 Esophageal Adenocarcinomas Defines Genomic Biomarkers for the Clinic'. In: *Nature Genetics* 51.3 (Mar. 2019), pp. 506–516. (Visited on 22/06/2023).

[224] Maria Secrier et al. 'Mutational Signatures in Esophageal Adenocarcinoma Define Etiologically Distinct Subgroups with Therapeutic Relevance'. In: *Nature Genetics* 48.10 (Oct. 2016), pp. 1131–1141. (Visited on 22/06/2023).

[225] P. C. Galipeau et al. 'Clonal Expansion and Loss of Heterozygosity at Chromosomes 9p and 17p in Premalignant Esophageal (Barrett's) Tissue'. In: *Journal of the National Cancer Institute* 91.24 (Dec. 1999), pp. 2087–2095.

[226] Carlo C. Maley et al. 'Genetic Clonal Diversity Predicts Progression to Esophageal Adenocarcinoma'. In: *Nature Genetics* 38.4 (Apr. 2006), pp. 468–473. (Visited on 14/10/2023).

[227] Sarah Killcoyne and Rebecca C. Fitzgerald. 'Evolution and Progression of Barrett's Oesophagus to Oesophageal Cancer'. In: *Nature Reviews Cancer* 21.11 (Nov. 2021), pp. 731–741. (Visited on 22/06/2023).

[228] Thomas Helleday, Saeed Eshtad and Serena Nik-Zainal. 'Mechanisms Underlying Mutational Signatures in Human Cancers'. In: *Nature Reviews Genetics* 15.9 (Sept. 2014), pp. 585–598. (Visited on 15/10/2023).

[229] Austin M. Dulak et al. 'Exome and Whole-Genome Sequencing of Esophageal Adenocarcinoma Identifies Recurrent Driver Events and Mutational Complexity'. In: *Nature Genetics* 45.5 (May 2013), pp. 478–486. (Visited on 15/10/2023).

[230] Sujath Abbas et al. 'Mutational Signature Dynamics Shaping the Evolution of Oesophageal Adenocarcinoma'. In: *Nature Communications* 14.1 (July 2023), p. 4239. (Visited on 16/09/2023).

[231] Lutz Krause et al. 'Identification of the CIMP-like Subtype and Aberrant Methylation of Members of the Chromosomal Segregation and Spindle Assembly Pathways in Esophageal Adenocarcinoma'. In: *Carcinogenesis* 37.4 (Apr. 2016), pp. 356–365. (Visited on 13/07/2023).

[232] Muhammad A. Alvi et al. 'DNA Methylation as an Adjunct to Histopathology to Detect Prevalent, Inconspicuous Dysplasia and Early-Stage Neoplasia in Barrett's Esophagus'. In: *Clinical Cancer Research* 19.4 (Feb. 2013), pp. 878–888. (Visited on 15/10/2023).

[233] E. Georg Luebeck et al. 'Identification of a Key Role of Widespread Epigenetic Drift in Barrett's Esophagus and Esophageal Adenocarcinoma'. In: *Clinical Epigenetics* 9 (Oct. 2017), p. 113. (Visited on 26/06/2023).

[234] Ming Yu et al. 'Subtypes of Barrett's Oesophagus and Oesophageal Adenocarcinoma Based on Genome-Wide Methylation Analysis'. In: *Gut* 68.3 (Mar. 2019), pp. 389–399. (Visited on 22/06/2023).

[235] Yang Liu et al. 'Comparative Molecular Analysis of Gastrointestinal Adenocarcinomas'. In: *Cancer Cell* 33.4 (Apr. 2018), 721–735.e8. (Visited on 22/06/2023).

[236] Yueyuan Zheng et al. 'Comprehensive Analyses of Partially Methylated Domains and Differentially Methylated Regions in Esophageal Cancer Reveal Both Cell-Type- and Cancer-Specific Epigenetic Regulation'. In: *Genome Biology* 24 (Aug. 2023), p. 193. (Visited on 16/10/2023).

[237] Thomas M. Carroll et al. 'Tumor Monocyte Content Predicts Immunochemotherapy Outcomes in Esophageal Adenocarcinoma'. In: *Cancer Cell* 41.7 (July 2023), 1222–1241.e7.

[238] Oliver Bohnsack, Axel Hoos and Katarina Ludajic. 'Adaptation and Modification of the Immune Related Response Criteria (IRRC): IrRECIST.' In: *Journal of Clinical Oncology* 32.15_suppl (May 2014), e22121–e22121. (Visited on 17/10/2023).

[239] Lesley Seymour et al. 'iRECIST: Guidelines for Response Criteria for Use in Trials Testing Immunotherapeutics'. In: *The Lancet. Oncology* 18.3 (Mar. 2017), e143–e152. (Visited on 17/10/2023).

[240] Alexander Dobin et al. 'STAR: Ultrafast Universal RNA-seq Aligner'. In: *Bioinformatics* 29.1 (Jan. 2013), pp. 15–21. (Visited on 17/10/2023).

[241] Yang Liao, Gordon K. Smyth and Wei Shi. 'featureCounts: An Efficient General Purpose Program for Assigning Sequence Reads to Genomic Features'. In: *Bioinformatics* 30.7 (Apr. 2014), pp. 923–930. (Visited on 17/10/2023).

[242] Michael I. Love, Wolfgang Huber and Simon Anders. 'Moderated Estimation of Fold Change and Dispersion for RNA-seq Data with DESeq2'. In: *Genome Biology* 15.12 (Dec. 2014), p. 550. (Visited on 17/07/2023).

[243] Tinyi Chu et al. 'Cell Type and Gene Expression Deconvolution with BayesPrism Enables Bayesian Integrative Analysis across Bulk and Single-Cell RNA Sequencing in Oncology'. In: *Nature Cancer* 3.4 (Apr. 2022), pp. 505–517. (Visited on 31/07/2023).

[244] Matthew Stephens. 'False Discovery Rates: A New Deal'. In: *Biostatistics* 18.2 (Apr. 2017), pp. 275–294. (Visited on 31/07/2023).

[245] Gennady Korotkevich et al. *Fast Gene Set Enrichment Analysis*. Feb. 2021. (Visited on 19/09/2023).

[246] Arthur Liberzon et al. 'The Molecular Signatures Database (MSigDB) Hallmark Gene Set Collection'. In: *Cell systems* 1.6 (Dec. 2015), pp. 417–425. (Visited on 19/09/2023).

[247] Heng Li and Richard Durbin. 'Fast and Accurate Short Read Alignment with Burrows–Wheeler Transform'. In: *Bioinformatics* 25.14 (July 2009), pp. 1754–1760. (Visited on 18/10/2023).

[248] Sangtae Kim et al. 'Strelka2: Fast and Accurate Calling of Germline and Somatic Variants'. In: *Nature Methods* 15.8 (Aug. 2018), pp. 591–594. (Visited on 18/10/2023).

[249] David Benjamin et al. *Calling Somatic SNVs and Indels with Mutect2*. Dec. 2019. (Visited on 18/10/2023).

[250] Daniel P. Cooke, David C. Wedge and Gerton Lunter. 'A Unified Haplotype-Based Method for Accurate and Comprehensive Variant Calling'. In: *Nature Biotechnology* 39.7 (July 2021), pp. 885–892. (Visited on 18/10/2023).

[251] Rachel Rosenthal et al. 'deconstructSigs: Delineating Mutational Processes in Single Tumors Distinguishes DNA Repair Deficiencies and Patterns of Carcinoma Evolution'. In: *Genome Biology* 17.1 (Feb. 2016), p. 31. (Visited on 16/09/2023).

[252] Serena Nik-Zainal et al. 'The Life History of 21 Breast Cancers'. In: *Cell* 149.5 (May 2012), pp. 994–1007. (Visited on 14/07/2023).

[253] Stefan C. Dentro, David C. Wedge and Peter Van Loo. 'Principles of Reconstructing the Subclonal Architecture of Cancers'. In: *Cold Spring Harbor Perspectives in Medicine* 7.8 (Aug. 2017), a026625. (Visited on 14/07/2023).

[254] Alice Antonello et al. *Computational Validation of Clonal and Subclonal Copy Number Alterations from Bulk Tumour Sequencing*. May 2023. (Visited on 18/10/2023).

[255] Stefan C. Dentro et al. 'Characterizing Genetic Intra-Tumor Heterogeneity across 2,658 Human Cancer Genomes'. In: *Cell* 184.8 (Apr. 2021), 2239–2254.e39. (Visited on 20/09/2023).

[256] Peng Jia et al. 'MSIsensor-pro: Fast, Accurate, and Matched-normal-sample-free Detection of Microsatellite Instability'. In: *Genomics, Proteomics & Bioinformatics* 18.1 (Feb. 2020), pp. 65–71. (Visited on 18/10/2023).

[257] Yun Zhang et al. 'Rapid and Accurate Alignment of Nucleotide Conversion Sequencing Reads with HISAT-3N'. In: *Genome Research* (June 2021), gr.275193.120. (Visited on 18/10/2023).

[258] Devon Ryan. *MethylDackel*. Oct. 2023. (Visited on 18/10/2023).

[259] Yongseok Park and Hao Wu. 'Differential Methylation Analysis for BS-seq Data under General Experimental Design'. In: *Bioinformatics* 32.10 (May 2016), pp. 1446–1453. (Visited on 14/07/2023).

[260] Yufang Qin et al. 'InfiniumPurify: An R Package for Estimating and Accounting for Tumor Purity in Cancer Methylation Research'. In: *Genes & Diseases* 5.1 (Mar. 2018), pp. 43–45.

[261] Douglas Arneson, Xia Yang and Kai Wang. 'MethylResolver-a Method for Deconvoluting Bulk DNA Methylation Profiles into Known and Unknown Cell Contents'. In: *Communications Biology* 3.1 (Aug. 2020), p. 422.

[262] Bowen Liu et al. 'MEpurity: Estimating Tumor Purity Using DNA Methylation Data'. In: *Bioinformatics* 35.24 (Dec. 2019), pp. 5298–5300. (Visited on 17/07/2023).

[263] Mark D. Robinson et al. 'Copy-Number-Aware Differential Analysis of Quantitative DNA Sequencing Data'. In: *Genome Research* 22.12 (Dec. 2012), pp. 2489–2496. (Visited on 27/05/2021).

[264] Elizabeth Larose Cadieux et al. *Copy Number-Aware Deconvolution of Tumor-Normal DNA Methylation Profiles*. Apr. 2022. (Visited on 17/07/2023).

[265] Xiaoqi Zheng et al. 'MethylPurify: Tumor Purity Deconvolution and Differential Methylation Detection from Single Tumor DNA Methylomes'. In: *Genome Biology* 15.7 (Aug. 2014), p. 419. (Visited on 17/07/2023).

[266] James E. Barrett et al. 'Quantification of Tumour Evolution and Heterogeneity via Bayesian Epiallele Detection'. In: *BMC Bioinformatics* 18.1 (July 2017), p. 354. (Visited on 17/07/2023).

[267] Shicheng Guo et al. 'Identification of Methylation Haplotype Blocks Aids in Deconvolution of Heterogeneous Tissue Samples and Tumor Tissue-of-Origin Mapping from Plasma DNA'. In: *Nature Genetics* 49.4 (Apr. 2017), pp. 635–642. (Visited on 17/07/2023).

[268] Antti Häkkinen et al. 'Identifying Differentially Methylated Sites in Samples with Varying Tumor Purity'. In: *Bioinformatics* 34.18 (Sept. 2018), pp. 3078–3085. (Visited on 17/07/2023).

[269] Netanel Loyfer et al. 'A DNA Methylation Atlas of Normal Human Cell Types'. In: *Nature* 613.7943 (Jan. 2023), pp. 355–364. (Visited on 14/07/2023).

[270] Maxime Tarabichi et al. 'A Practical Guide to Cancer Subclonal Reconstruction from DNA Sequencing'. In: *Nature Methods* 18.2 (Feb. 2021), pp. 144–155. (Visited on 14/07/2023).

[271] Qiang Song et al. 'A Reference Methylome Database and Analysis Pipeline to Facilitate Integrative and Comparative Epigenomics'. In: *PLOS ONE* 8.12 (Dec. 2013), e81148. (Visited on 18/07/2023).

[272] Egor Dolzhenko and Andrew D. Smith. 'Using Beta-Binomial Regression for High-Precision Differential Methylation Analysis in Multifactor Whole-Genome Bisulfite Sequencing Experiments'. In: *BMC Bioinformatics* 15.1 (June 2014), p. 215. (Visited on 18/07/2023).

[273] Kasper D. Hansen, Benjamin Langmead and Rafael A. Irizarry. 'BSmooth: From Whole Genome Bisulfite Sequencing Reads to Differentially Methylated Regions'. In: *Genome Biology* 13.10 (Oct. 2012), R83. (Visited on 18/07/2023).

[274] Frank Jühling et al. 'Metilene: Fast and Sensitive Calling of Differentially Methylated Regions from Bisulfite Sequencing Data'. In: *Genome Research* (Dec. 2015). (Visited on 18/07/2023).

[275] Ha Vu and Jason Ernst. 'Universal Annotation of the Human Genome through Integration of over a Thousand Epigenomic Datasets'. In: *Genome Biology* 23.1 (Jan. 2022), p. 9. (Visited on 18/07/2023).

[276] R. Dennis Cook. 'Detection of Influential Observation in Linear Regression'. In: *Technometrics* 19.1 (1977), pp. 15–18. JSTOR: 1268249. (Visited on 17/07/2023).

[277] Yoav Benjamini and Yosef Hochberg. 'Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing'. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 57.1 (1995), pp. 289–300. JSTOR: 2346101. (Visited on 20/07/2023).

[278] Arie B. Brinkman et al. 'Partially Methylated Domains Are Hypervariable in Breast Cancer and Fuel Widespread CpG Island Hypermethylation'. In: *Nature Communications* 10.1 (Apr. 2019), p. 1749. (Visited on 25/07/2023).

[279] Jianfeng Xu et al. 'Cellular Heterogeneity–Adjusted cLonal Methylation (CHALM) Improves Prediction of Gene Expression'. In: *Nature Communications* 12.1 (Jan. 2021), p. 400. (Visited on 29/08/2023).

[280] Jiejun Shi et al. 'The Concurrence of DNA Methylation and Demethylation Is Associated with Transcription Regulation'. In: *Nature Communications* 12.1 (Sept. 2021), p. 5285. (Visited on 29/08/2023).

[281] Melanie Ehrlich. 'DNA Hypomethylation in Cancer Cells'. In: *Epigenomics* 1.2 (Dec. 2009), pp. 239–259. (Visited on 29/08/2023).

[282] Venkatraman E. Seshan and Adam Olshen. *DNAcopy: DNA Copy Number Data Analysis*. Manual. 2020.

[283] Marjan M. Naeini et al. 'Multi-Omic Features of Oesophageal Adenocarcinoma in Patients Treated with Preoperative Neoadjuvant Therapy'. In: *Nature Communications* 14 (May 2023), p. 3155. (Visited on 22/06/2023).

[284] Weiwei Zhang et al. 'Accounting for Tumor Purity Improves Cancer Subtype Classification from DNA Methylation Data'. In: *Bioinformatics* 33.17 (Sept. 2017), pp. 2651–2657. (Visited on 07/08/2023).

[285] Jamie L. Endicott et al. 'Cell Division Drives DNA Methylation Loss in Late-Replicating Domains in Primary Human Cells'. In: *Nature Communications* 13.1 (Nov. 2022), p. 6659. (Visited on 25/07/2023).

[286] Nicholas Rhind and David M. Gilbert. 'DNA Replication Timing'. In: *Cold Spring Harbor Perspectives in Biology* 5.8 (Aug. 2013), a010132. (Visited on 25/07/2023).

[287] Benjamin E. Decato et al. 'Characterization of Universal Features of Partially Methylated Domains across Tissues and Species'. In: *Epigenetics & Chromatin* 13.1 (Oct. 2020), p. 39. (Visited on 11/01/2022).

[288] Jesse R. Dixon et al. 'Integrative Detection and Analysis of Structural Variation in Cancer Genomes'. In: *Nature Genetics* 50.10 (Oct. 2018), pp. 1388–1398. (Visited on 25/07/2023).

[289] Xena Giada Pappalardo and Viviana Barra. 'Losing DNA Methylation at Repetitive Elements and Breaking Bad'. In: *Epigenetics & Chromatin* 14.1 (June 2021), p. 25. (Visited on 26/07/2023).

[290] Amir Eden et al. 'Chromosomal Instability and Tumors Promoted by DNA Hypomethylation'. In: *Science* 300.5618 (Apr. 2003), pp. 455–455. (Visited on 26/07/2023).

[291] G. Howard et al. 'Activation and Transposition of Endogenous Retroviral Elements in Hypomethylation Induced Tumors in Mice'. In: *Oncogene* 27.3 (Jan. 2008), pp. 404–408. (Visited on 26/07/2023).

[292] Hironobu Shigaki et al. 'LINE-1 Hypomethylation in Gastric Cancer, Detected by Bisulfite Pyrosequencing, Is Associated with Poor Prognosis'. In: *Gastric Cancer* 16.4 (2013), pp. 480–487. (Visited on 26/07/2023).

[293] Yu Kong et al. 'Transposable Element Expression in Tumors Is Associated with Immune Infiltration and Increased Antigenicity'. In: *Nature Communications* 10.1 (Nov. 2019), p. 5228. (Visited on 26/07/2023).

[294] Anshul Kundaje et al. 'Integrative Analysis of 111 Reference Human Epigenomes'. In: *Nature* 518.7539 (Feb. 2015), pp. 317–330. (Visited on 30/07/2023).

[295] Rory T. Coleman and Gary Struhl. 'Causal Role for Inheritance of H3K27me3 in Maintaining the OFF State of a Drosophila HOX Gene'. In: *Science* 356.6333 (Apr. 2017), eaai8236. (Visited on 28/07/2023).

[296] Friederike Laprell, Katja Finkl and Jürg Müller. 'Propagation of Polycomb-repressed Chromatin Requires Sequence-Specific Recruitment to DNA'. In: *Science* 356.6333 (Apr. 2017), pp. 85–88. (Visited on 28/07/2023).

[297] Emmanuelle Viré et al. 'The Polycomb Group Protein EZH2 Directly Controls DNA Methylation'. In: *Nature* 439.7078 (Feb. 2006), pp. 871–874. (Visited on 28/07/2023).

[298] Francesco Neri et al. 'Genome-Wide Analysis Identifies a Functional Association of Tet1 and Polycomb Repressive Complex 2 in Mouse Embryonic Stem Cells'. In: *Genome Biology* 14.8 (Aug. 2013), R91. (Visited on 28/07/2023).

[299] Yuanyuan Li et al. 'Genome-Wide Analyses Reveal a Role of Polycomb in Promoting Hypomethylation of DNA Methylation Valleys'. In: *Genome Biology* 19.1 (Feb. 2018), p. 18. (Visited on 14/07/2023).

[300] Raha Weigert et al. 'Dynamic Antagonism between Key Repressive Pathways Maintains the Placental Epigenome'. In: *Nature Cell Biology* 25.4 (Apr. 2023), pp. 579–591. (Visited on 28/07/2023).

[301] James P. Reddington et al. 'Redistribution of H3K27me3 upon DNA Hypomethylation Results in De-Repression of Polycomb Target Genes'. In: *Genome Biology* 14.3 (Mar. 2013), R25. (Visited on 28/07/2023).

[302] Yuta Takahashi et al. 'Integration of CpG-free DNA Induces de Novo Methylation of CpG Islands in Pluripotent Stem Cells'. In: *Science* 356.6337 (May 2017), pp. 503–508. (Visited on 29/07/2023).

[303] Yang Lu et al. 'Pan-Cancer Analysis Revealed H3K4me1 at Bivalent Promoters Premarks DNA Hypermethylation during Tumor Development and Identified the Regulatory Role of DNA Methylation in Relation to Histone Modifications'. In: *BMC Genomics* 24.1 (May 2023), p. 235. (Visited on 30/07/2023).

[304] Simon Fishilevich et al. 'GeneHancer: Genome-Wide Integration of Enhancers and Target Genes in GeneCards'. In: *Database: The Journal of Biological Databases and Curation* 2017 (Apr. 2017), bax028. (Visited on 30/07/2023).

[305] Ran Zhao et al. 'Implications of Genetic and Epigenetic Alterations of CDKN2A (p16INK4a) in Cancer'. In: *eBioMedicine* 8 (June 2016), pp. 30–39. (Visited on 23/09/2023).

[306] Jonathan B. Weitzman. 'p16Ink4a and p19Arf: Terrible Twins'. In: *Trends in Molecular Medicine* 7.11 (Nov. 2001), p. 489. (Visited on 16/10/2023).

[307] Laura J. Hardie et al. 'P16 Expression in Barrett's Esophagus and Esophageal Adenocarcinoma: Association with Genetic and Epigenetic Alterations'. In: *Cancer Letters* 217.2 (Jan. 2005), pp. 221–230. (Visited on 16/10/2023).

[308] Viktor A. Adalsteinsson et al. 'Scalable Whole-Exome Sequencing of Cell-Free DNA Reveals High Concordance with Metastatic Tumors'. In: *Nature Communications* 8.1 (Nov. 2017), p. 1324. (Visited on 22/12/2022).

[309] Rafal Dziadziuszko et al. 'Circulating Cell-free DNA as a Prognostic Biomarker in Patients with Advanced ALK+ Non–Small Cell Lung Cancer in the Global Phase III ALEX Trial'. In: *Clinical Cancer Research* 28.9 (May 2022), pp. 1800–1808. (Visited on 25/04/2023).

[310] Emmalyn Chen et al. 'Cell-Free DNA Concentration and Fragment Size as a Biomarker for Prostate Cancer'. In: *Scientific Reports* 11.1 (Mar. 2021), p. 5040. (Visited on 25/04/2023).

[311] Daniel Fernandez-Garcia et al. 'Plasma Cell-Free DNA (cfDNA) as a Predictive and Prognostic Marker in Patients with Metastatic Breast Cancer'. In: *Breast Cancer Research* 21.1 (Dec. 2019), p. 149. (Visited on 25/04/2023).

[312] Edith Borcoman et al. 'Patterns of Response and Progression to Immunotherapy'. In: *American Society of Clinical Oncology Educational Book* 38 (May 2018), pp. 169–178. (Visited on 09/11/2023).

[313] Duncan Sproul and Richard R. Meehan. 'Genomic Insights into Cancer-Associated Aberrant CpG Island Hypermethylation'. In: *Briefings in Functional Genomics* 12.3 (May 2013), pp. 174–190. (Visited on 08/09/2023).

[314] Katherine A. Hoadley et al. 'Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer'. In: *Cell* 173.2 (Apr. 2018), 291–304.e6. (Visited on 18/09/2023).

[315] Savitri Krishnamurthy and Yogeshwar Dayal. 'Pancreatic Metaplasia in Barrett's Esophagus An Immunohistochemical Study'. In: *The American Journal of Surgical Pathology* 19.10 (Oct. 1995), p. 1172. (Visited on 18/09/2023).

[316] Nicolas A. Villa et al. 'Pancreatic Acinar Metaplasia at the Gastroesophageal Junction: A Single Institution Experience: 1730'. In: *Official journal of the American College of Gastroenterology | ACG* 110 (Oct. 2015), S735. (Visited on 18/09/2023).

[317] Takafumi Fuchino et al. 'Clinicopathological Characteristics of Pancreatic Acinar Cell Metaplasia Associated with Helicobacter Pylori Infection'. In: *BMC Gastroenterology* 22.1 (June 2022), p. 289. (Visited on 18/09/2023).

[318] Kevin Litchfield et al. 'Meta-Analysis of Tumor- and T Cell-Intrinsic Mechanisms of Sensitization to Checkpoint Inhibition'. In: *Cell* 184.3 (Feb. 2021), 596–614.e14. (Visited on 16/09/2023).

[319] Linde M. Veen et al. 'The Role of Transforming Growth Factor β in Upper Gastrointestinal Cancers: A Systematic Review'. In: *Cancer Treatment Reviews* 100 (Nov. 2021), p. 102285. (Visited on 19/09/2023).

[320] Qi Wang et al. 'Gene Body Methylation in Cancer: Molecular Mechanisms and Clinical Applications'. In: *Clinical Epigenetics* 14.1 (Nov. 2022), p. 154. (Visited on 19/09/2023).

[321] Dvir Aran, Sivan Sabato and Asaf Hellman. 'DNA Methylation of Distal Regulatory Sites Characterizes Dysregulation of Cancer Genes'. In: *Genome Biology* 14.3 (Mar. 2013), R21. (Visited on 10/10/2023).

[322] Dvir Aran, Marina Sirota and Atul J. Butte. 'Systematic Pan-Cancer Analysis of Tumour Purity'. In: *Nature Communications* 6.1 (Dec. 2015), p. 8971. (Visited on 18/10/2023).

[323] Yon Hee Kim et al. 'Histologic Purity of Signet Ring Cell Carcinoma Is a Favorable Risk Factor for Lymph Node Metastasis in Poorly Cohesive, Submucosa-Invasive Early Gastric Carcinoma'. In: *Gastric Cancer: Official Journal of the International Gastric Cancer Association and the Japanese Gastric Cancer Association* 20.4 (July 2017), pp. 583–590.