Exploring cell-type-specific cytosine modifications in somatic tissues and cancer

Magdalena Drożdż

Oriel College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

2023

Abstract

Cytosine methylation (5mC) is one of the most widely studied DNA modifications and plays crucial roles in correct development, regulation of gene expression, and maintaining genome stability. 5mC landscapes vary across human cell types and can be significantly altered in cancers. Differences in methylation patterns are frequently used in studying human development and disease and can be used as prognostic biomarkers. In this thesis, I show a novel, semi-supervised method based on non-negative matrix factorisation (NMF) for the identification of highly variable, tissue-specific CpG blocks across the genome. The method is based on a TAPSβ atlas representing a collection of 20 different cells and tissues. I applied the method to both heterogeneous data and FACS-sorted cells, obtaining satisfying results in cross-validation, finding that the method required minimal optimisation to be applicable to new datasets. The selected blocks cover a wide range of genomic contexts and are co-localised with tissue-specific genes, histone marks, chromatin states, and are enriched in 5hmC signals across most tissues. I present how the method can be used in the cancer context, by deconvoluting samples from patients with oesophageal adenocarcinoma (OAC) enroled in the LUD2015-005 clinical trial. The results gave insights into the cellular composition of patient samples and tumour purity and found that changes in the contribution of stomach and eosinophils between time points were associated with different survival probabilities.

Exploring cell-type-specific cytosine modifications in somatic tissues and cancer

Magdalena Drożdż

Oriel College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

2023

This thesis is dedicated to my grandfather, Bohdan (1926–2009).
Though time has not allowed our discussions on science,
perhaps my fascination with DNA is simply hereditary.

Acknowledgements

The four years spent doing my DPhil were a true roller coaster. It had everything: exciting highs, dramatic lows, unexpected turns, and a constant thrill of novelty at each step. I would like to thank my supervisor Prof. Benjamin Schuster-Boeckler for offering me this opportunity, for supporting my academic progress, and for always believing in me a little bit more than I did. I would also like to thank Prof. Skirmantas Kriaucionis for his share of the supervision – our meetings always left me with new sparks of excitement about science, even in times where all I wanted to do was to open a cafe and never read a single academic paper ever again.

I would like to acknowledge the entire Ludwig Institute for Cancer Research team. It has been a fantastic place to grow as a scientist and I am very grateful for the learning opportunities and support I received during my time here. I would also like to thank Oriel College, which welcomed me so warmly and recognised my work by granting me the Oriel Graduate Scholarship.

The Oxford experience would not be the same if it were not for the amazing people I had the opportunity to meet here. I am afraid there would not be enough paper in the entire city to print this thesis if I listed everyone I am grateful for, so I will try my best to keep it to a reasonable length.

First, I would like to thank the entire SK lab and the lunch club – Michael, Emelie, Felice, Olly, Tom, Lucy, Ashvina, Nandini, and Matthew. I will never forget our conversations over lunch and in the Vitamin D Deficiency Corner, and your support in the difficult times, in the form of memes, sweets, or an ungodly amount of cheese (thankfully, the last one only happened once). I am also thankful to all the BSB lab members - the old, the new, and the short-term ones. It was a pleasure doing science with you, and believe me, it is not normally that easy to meet people who have sparks of excitement in their eyes when they talk about some obscure R packages (looking at you, Phil!). This is also a good place to mention Olena - thank you and Mike for being the best adoptive Oxford parents, providing fantastic food, career advice, dog time and large amounts of gin and tonic.

I would also like to thank everyone at the Oxford University Mountaineering Club - I believe that the friendships (and injuries) I made at the club will last for many

long years. I am grateful for all the wonderful memories I made with you – from the sunny cliffs of Spain, to sweaty bouldering rooms, to cold, rainy rocks of the Lake District, sometimes so miserable that they made working on the DPhil look exceptionally fun and cosy in comparison. Again, I cannot possibly acknowledge everyone individually, but a few notable mentions include Eddy, Maisie, Giorgia, Nadav and Faiz – you made this sport mean so much more than just crimping on plastic holds (or, weather permitting, real rocks).

Next, I would like to thank the Polish contingent. Working in the Polonium Foundation along with the DPhil has been incredibly fun and fulfilling, and I will be forever grateful for meeting this bunch of extremely talented and hard-working group of volunteers. I would like to especially mention Nika, Kasia, Alicja and Tomek, whose friendship and support meant so much – you always believed in me and pushed me to do things that I would not otherwise have the courage to do.

I am eternally grateful for everyone back at home – my family supported me unconditionally, despite some truly difficult times that coincided with the DPhil. I would especially like to thank my amazing big brother Michał, who was indispensable in the last few months and managed to keep me grounded even when my thoughts and fears were flying in what seemed to be all directions, at the same time, at full speed. I am grateful to my mum Iwona, for always supporting my choices and letting me follow my (sometimes questionable) ideas and dreams. This also goes to my dad, Grzegorz – although stomach cancer took you from us nine years ago, I know that you have been watching over me all this time. Having half of my DNA coming from you *really* showed during the DPhil (perhaps for an added extra challenge).

And last but not least, a huge acknowledgement to my best friend, Monika. I cannot thank you enough. You have been everything a friend could ever wish for.

Abstract

Cytosine methylation (5mC) is one of the most widely studied DNA modifications and plays crucial roles in correct development, regulation of gene expression, and maintaining genome stability. 5mC landscapes vary across human cell types and can be significantly altered in cancers. Differences in methylation patterns are frequently used in studying human development and disease and can be used as prognostic biomarkers. In this thesis, I show a novel, semi-supervised method based on non-negative matrix factorisation (NMF) for the identification of highly variable, tissue-specific CpG blocks across the genome. The method is based on a TAPSβ atlas representing a collection of 20 different cells and tissues. I applied the method to both heterogeneous data and FACS-sorted cells, obtaining satisfying results in cross-validation, finding that the method required minimal optimisation to be applicable to new datasets. The selected blocks cover a wide range of genomic contexts and are co-localised with tissue-specific genes, histone marks, chromatin states, and are enriched in 5hmC signals across most tissues. I present how the method can be used in the cancer context, by deconvoluting samples from patients with oesophageal adenocarcinoma (OAC) enroled in the LUD2015-005 clinical trial. The results gave insights into the cellular composition of patient samples and tumour purity and found that changes in the contribution of stomach and eosinophils between time points were associated with different survival probabilities.

Contents

Appendices

List of Abbreviations

3' UTR . . . . . 3' Untranslated Region

5' UTR . . . . . 5' Untranslated Region

5caC 5-carboxylcytosine

5fC 5-formylcytosine

5hmC 5-hydroxymethylcytosine

5mC 5-methylcytosine

BER . . . . . Base Excision Repair

bp . . . . . Base pair

CGI . . . . . CpG island

ChIP-Seq . . . . Chromatin Immunoprecipitation Sequencing

CNS . . . . . Central Nervous System

CTX . . . . . Chemotherapy

DMR . . . . . Differentially Methylated Region

DNMT . . . . . DNA Methyltransferase

FACS . . . . . Fluorescence-activated Cell Sorting

GEJ . . . . . Gastroesophageal junction

GO . . . . . Gene Ontology

GO BP . . . . . Gene Ontology Biological Process

HAT . . . . . Histone Acetyltransferase

HCP . . . . . High CpG Promoter

ICI . . . . . Immune Checkpoint Inhibitor

KEGG . . . . . Kyoto Encyclopedia of Genes and Genomes

LCP . . . . . Low CpG Promoter

MANE . . . . . Matched Annotation from the NCBI and EMBL-EBI

MBD . . . . . Methyl-CpG Binding Domain

MBD-Seq . . . Methyl-CpG Binding Domain Sequencing

MeDIP-Seq . . . Methylated DNA Immunoprecipitation Sequencing

NK cell . . . . . Natural Killer Cell

NGS . . . . . Next Generation Sequencing

NNLS . . . . . Non-Negative Least Squares

NMF . . . . . Non-negative Matrix Factorization

OAC . . . . . Oesophageal Adenocarcinoma

ORA . . . . . Overrepresentation Analysis

OS . . . . . Overall Survival

OSCC . . . . . Oesophageal Squamous Cell Carcinoma

PCA . . . . . Principal Component Analysis

PCR . . . . . Polymerase Chain Reaction

PRC2 . . . . . Polycomb Repressive Complex 2

RNAP . . . . . RNA Polymerase

SNP . . . . . Single Nucleotide Polymorphism

SNV . . . . . Single Nucleotide Variant

TAB-SEQ . . . . TET-assisted Bisulfite Sequencing

TAPS . . . . . TET-assisted pyridine borane sequencing

TDG . . . . . Thymine DNA Glycosylase

TET . . . . . Ten-Eleven Translocation (Protein)

TF . . . . . Transcription Factor

TFBS . . . . . Transcription Factor Binding Site

THPA . . . . . The Human Protein Atlas

TSS . . . . . Transcription Start Site

1 Introduction

Contents

1.1 Introduction to epigenomics

The development of a multicellular organism is a captivating phenomenon. It begins with a single cell, which contains a unique combination of maternal and paternal DNA, which gradually unfolds into a complex organism composed of a myriad of cells, tissues, and organs. Each constituent plays a unique role, but acts together in the orchestra, underlying the intricate balance of life. Intriguingly, the individual cells comprising these larger structures share identical DNA, or *genomes* – the same set of "building instructions" derived from the original cell. However, during development, different parts of the instructions are interpreted

by different cells, leading to cellular differentiation. This process equips cells with distinct structures, functionalities, and characteristics: from melanin production in melanocytes, to bile acid synthesis in liver cells, and light perception in the eye. Despite sharing the same DNA, each cell manifests a different gene expression profile, which remains malleable through development, and responsive to environmental changes. Various molecular systems are responsible for establishing these differences, and one of them is called epigenetics (from Greek *-epi*, "above" genetics).

This term was coined by Conrad Waddington in the early 1940s to describe the study of causal interactions between genes and phenotypes (H 1942). During the rest of the century, the term evolved to encompass the role of epigenetics in development, its heritability, mediation by the environment, and the discoveries of new 'epigenetic' factors (Jablonka and Lamb 2002; Haig 2004). As a result, the definition comes in slightly different flavours depending on the field of study (Haig 2004; Deans and Maggert 2015). In the context of genomics, the most widespread definition is based on the work of Holliday and later modified by Wu and Morris: "the study of changes in gene function that are mitotically and/or meiotically heritable and that do not entail change in DNA sequence" (Wu Ct and Morris 2001; Holliday 1994). In this thesis, I will focus on modifications of the DNA bases and proteins interacting with DNA, and the term "epigenomics" will be used to describe the collection of all epigenetic changes in the genome.

The human genome is composed of 3.1 billion base pairs (Gbp), made up of adenine (A), cytosine (C), guanine (G), and thymine (T), linked in a chain-like sequence (Nurk et al. 2022). This chain is divided into 23 chromosomes. Its accessibility and functionality are regulated by a higher-order structure known as chromatin (Felsenfeld and Groudine 2003). Chromatin is mainly made up of nucleosomes, which are repeating units of 147 base pairs of DNA wrapped around a histone octamer containing two copies each of the core histones: H2A, H2B, H3, and H4 (McGinty and Tan 2015). Under electron microscopy, the periodic arrangement of nucleosomes gives chromatin a 'beads-on-a-string' appearance. The nucleosome positions along with histone variants and modifications are essential in modulating the chromatin structure and accessibility of various proteins and transcription factors, subsequently affecting gene expression (Bernstein et al. 2007; Margueron and Reinberg 2010). Histone proteins are subject to a myriad of post-translational modifications, including but not limited to acetylation, methylation and phosphorylation. Each modification, depending on its type, position, and context, may be distinctly associated with the functional outcomes of the associated

DNA – for example, H3 lysine 4 trimethylation (H3K4me3) is strongly enriched in active promoters, while H3 lysine 27 trimethylation (H3K27me3) is associated with repressed states (Wu and Zhang 2015). Mapping histone modifications and related structures using chromatin immunoprecipitation followed by sequencing (ChIP–seq) on a genome-wide scale has become a powerful tool for determining the factors that influence chromatin structure, as well as for comprehending the part that epigenetic modifications play in gene regulation.

1.2 DNA methylation

Cytosine methylation is the most extensively studied form of epigenetic DNA modification. Initially described as a heritable epigenetic mark in 1975 (Holliday and Pugh 1975; Riggs 1975), this process involves the covalent attachment of a methyl group to the 5-carbon position of the cytosine ring within the DNA, forming 5-methylcytosine (5mC) (Bird and Taggart 1980; Bestor 2000). Evident in numerous organisms, methylation has been identified in several species of bacteria, fungi, plants, invertebrates, and all investigated vertebrates (Suzuki and Bird 2008; Capuano et al. 2014; Su et al. 2011). However, it remains absent in *Caenorhabditis elegans*, insects, and yeast strains studied to date (Capuano et al. 2014; Simpson et al. 1986; Varma et al. 2022). Despite the ubiquity of this chemical modification, the genomic context and the presumed function vary significantly between species (Suzuki and Bird 2008; Varma et al. 2022). The following section will exclusively focus on methylation as observed in mammals, especially humans.

Approximately 4% of all cytosines in the human genome are methylated, accounting for ̃1% of all bases. Most DNA methylation occurs at CpG dinucleotides (CpG sites). Of the 32.8 million CpG pairs in the genome, around 75% are methylated (Gershman et al. 2022). Methylation patterns exhibit variation between cell types and are associated with differential gene expression programmes in various tissues (Jones and Takai 2001). Non-CpG methylation is rare in most somatic cells, but has been detected in the mCHG and mCHH contexts (where H = A, C, or T) in cells such as neurons and embryonic stem cells, respectively (Lister et al. 2009; He and Ecker 2015).

Early studies indicated the role of 5mC in transcriptional repression (Ben-Hattar and Jiricny 1988; Watt and Molloy 1988; Iguchi-Ariga and Schaffner 1989), and in larger gene silencing phenomena such as genomic imprinting (Ferguson-Smith et al. 1993; Li et al. 1993) and X inactivation (Jaenisch and Bird 2003).

Methylation has since been found to play a role in numerous biological processes, from early development to cellular differentiation and regulation of pluripotency to maintenance of genome stability (relevant areas will be described in more detail in the following sections) (Jones 2012). However, "With great power comes great responsibility", and indeed, abnormalities in 5mC landscapes are frequently linked to disease and disordered development (Lee and Ditko 1962).

1.2.1 Methylation writers and erasers

The process of depositing methylation on DNA is catalysed by a family of enzymes known as DNA methyltransferases (DNMTs) (Figure 1.1). In humans, several DNMTs are active – DNMT1, DNMT3A, and DNMT3B (Yoder and Bestor 1998; Okano et al. 1999). DNMT1 is referred to as the *maintenance methyltransferase* as it primarily deposits methylation on hemimethylated DNA during DNA replication, aided by UHRF1 (Bostick et al. 2007). DNMT3A and DNMT3B, commonly known as *de novo methyltransferases*, establish new methylation marks during development and cell differentiation by methylating previously unmethylated CpG sites (reviewed in Law and Jacobsen 2010 and Moore et al. 2013). The presence and correct functioning of all these methyltransferases are essential for normal development. For example, complete knockout of *Dnmt1* in mice results in embryonic lethality (Li et al. 1992). Moreover, insufficiency or mutations in any of these can lead to severe developmental defects, such as ICF syndrome which is specifically caused by DNMT3B abnormalities (Okano et al. 1999; Hirasawa et al. 2008; Xu et al. 1999; Klein et al. 2013).

Methylation can be lost passively through imperfect maintenance (replication without the presence of DNMT1), or actively, aided by the ten-eleven translocation (TET) family of proteins (Tahiliani et al. 2009). TET proteins oxidise 5mC to 5-hydroxymethylcytosine (5hmC) and further to 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC) (He et al. 2011; Ito et al. 2011). All of these modifications can lead to the dilution of methylation after DNA replication in the process of passive demethylation. Alternatively, 5caC and 5fC can be targeted by a base excision repair (BER) enzyme thymine DNA glycosylase (TDG), leading to their excision and replacement with an unmodified cytosine in the replication-independent demethylation pathway (Maiti and Drohat 2011; He et al. 2011).

The activity of the TET proteins has been detected in multiple processes that span development, meiosis, maintenance of imprinting, cell reprogramming, and transcription (Gu et al. 2011; Quivoron et al. 2011; Ficz et al. 2011; Dawlaty et al.

2013; Wu et al. 2011b; Ginno et al. 2020), and transient accumulation of 5caC has been found in lineage specification (Wheldon et al. 2014). Despite the active conversion by TET proteins, some of these modifications have been found to be stable in the genome (Bachman et al. 2014).

The brain and central nervous system (CNS) are particularly rich in 5hmC, which emphasises its role in the development and maintenance of neural tissue (Kriaucionis and Heintz 2009; Globisch et al. 2010; Guo et al. 2011; Szulwach et al. 2011; He et al. 2021). Although detected at lower levels in multiple other tissues, the presence of 5hmC is evident. 5fC and 5caC are much less abundant than 5hmC (10 to 1,000-fold) and are not enriched in any particular tissue (Ito et al. 2011). Given that they were found to mostly reside in sides of abundant 5hmC, they are considered to be committed demethylation intermediates and mark regions undergoing active demethylation (Song et al. 2012).

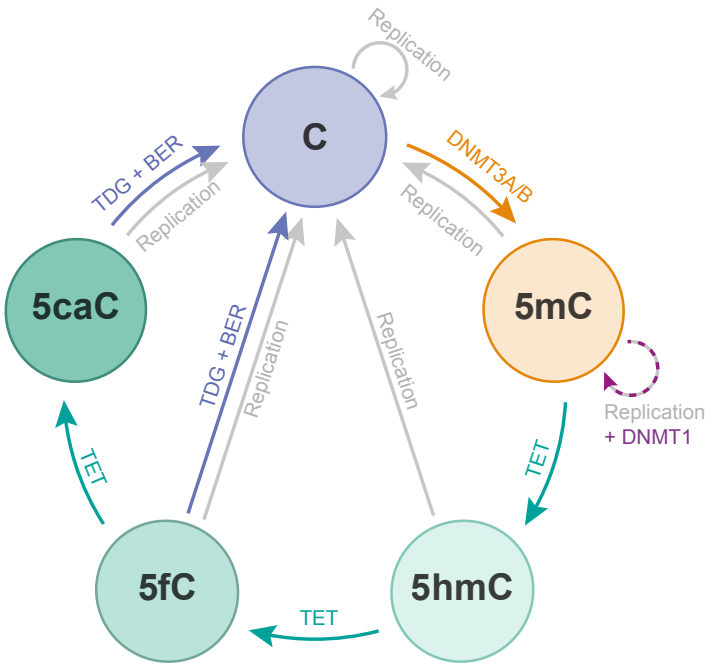


Figure 1.1: Passive and active demethylation pathways
Cytosine can become methylated *de novo* in a reaction catalysed by DNMT3A / B, or after DNA replication in the presence of DNMT1. Passive demethylation occurs as a result of DNA replication without the activity of DNMT1. TET-assisted passive demethylation happens after replication of DNA with products of TET oxidation. TET-assisted active demethylation occurs after 5fC or 5caC are excised from DNA and replaced by unmodified cytosine during base excision repair. All cytosines in this figure are in the CpG context. Figure based on Raiber et al. 2017.

1.2.2 Techniques to study cytosine modifications

The field of epigenomics has been closely reliant on the development of DNA sequencing technologies. The challenge in detecting a chemical modification as small as a methyl group led to the development of a wide array of methods, ranging from gel-, array- and sequencing-based methods which differ in their resolution, sensitivity, costs, and applications (extensively reviewed in Beck and Rakyan 2008; Wu and Zhang 2015).

One class of methylation detection techniques is affinity-based, including methods such as methyl-binding domain sequencing (MBD-Seq) and methyl-DNA immunoprecipitation sequencing (MeDIP-Seq). MBD-Seq targets MBD protein-bound regions, and MeDIP-seq uses 5mC-specific monoclonal antibodies to enrich methylated DNA fragments. DNA is sequenced after enrichment, and depending on the method used, the result represents 17. 8% of CpG sites in the case of MBD-Seq and 87.5% in MeDIP-seq (when coupled with Next Generation Sequencing, NGS) (Stirzaker et al. 2014; Clark et al. 2012). Furthermore, these methods have been adjusted to detect other cytosine modifications by, for example, using 5hmC-specific antibodies in hMeDIP-seq. A range of other methods has been developed to study 5hmC, including CMS-seq, GLIB, and hME-SEAL, all based on enrichment followed by sequencing (Wu and Zhang 2015). The problem with enrichment-based methods is that they do not provide positional information and do not inform of the exact modification levels, as the genome is not covered evenly.

Methylation sequencing at single nucleotide resolution was made possible after the discovery of the properties of bisulfite treatment in 1973 (Shapiro et al. 1973). When DNA is treated with sodium bisulfite, unmethylated cytosines undergo a sulfonation reaction at their 4th position, followed by hydrolytic deamination, converting them into uracil-sulphonate, which is ultimately hydrolyzed to uracil. Methylated and hydroxymethylated cytosines, protected by their methyl group, remain unchanged during this process. During the subsequent PCR, the new uracils are paired with adenines instead of guanines. As a result, in the amplified DNA, all unmethylated cytosines are turned to thymines, while 5mC and 5hmC remain unchanged (Figure 1.2). This property is used in a variety of methylation sequencing techniques. Some of the most popular and cost-effective methods are array-based technologies, such as Illumina's HM450 array, which cover carefully selected CpGs in regulatory parts of the genome, accounting for 1.37% of all CpGs (Bibikova et al. 2011). Reduced representation bisulfite sequencing (RRBS)

which increases the CpG representation to 3.7%, by selecting CpG-dense regions using methylation-insensitive restriction enzymes, followed by NGS (Stirzaker et al. 2014). Although they offer positional information and exact methylation levels, these methods represent only a small fraction of the genome.

With the advent and subsequent cost reduction of whole genome sequencing (WGS), bisulfite began to be used to treat the DNA prior to sequencing. Whole genome bisulfite sequencing (WGBS) represents $\tilde{9}5\%$ of the genome (Stirzaker et al. 2014), providing a large-scale, single nucleotide resolution, and quantitative method to study DNA methylation. Due to the inability of WGBS to distinguish between 5hmC and 5mC, other methods followed: TET-assisted bisulfite sequencing (TAB-seq) which allows the identification of 5hmC and oxBS-seq, which reads exclusively 5mC. In theory, the measurements of TAB-seq and oxBS-seq can be combined for each position, and their difference can show levels of both modifications, but that has been proven to be challenging (Wu and Zhang 2015). While useful and incredibly popular, bisulfite conversion presents multiple challenges. It is known to be a harsh chemical treatment causing DNA damage and degradation, negatively impacting the sequencing results. Additionally, conversion of more than 95% of cytosines to thymines causes loss of sequence complexity, which complicates alignment of sequences to the reference genome during bioinformatic analysis, further causing issues with loss of unmappable or incorrectly mapped reads. (Clark et al. 2006; Pomraning et al. 2009)

A novel bisulfite-free method for studying cytosine modifications at base resolution is TET-assisted pyridine borane sequencing (TAPS) (Liu et al. 2019). In this technology, cytosine modifications (5mC, 5hmC, 5fC) are first oxidised to 5caC by the TET enzyme. This is followed by a treatment with pyridine borane, which reduces 5caC to 5,6-dihydrouracil (DHU). DHU, as a uracil derivative, is read as thymine during the subsequent PCR. This bisulfite-free method requires milder chemical conditions, and given the abundance of modified cytosines to all cytosines, only 1-5% of all cytosines will be converted, improving library complexity and read mapping. Moreover, the use of TAPS can be easily extended to cell-free DNA, in which the degradation followed by harsh bisulfite treatment is even more problematic due to low amounts of input DNA. Two additional methods have been developed to increase the modification detection sensitivity: TAPSβ, the bisulfite-free equivalent of oxBS-seq, which enables the identification of only 5mC due to the protection of 5hmC by β-glucosyltansferase treatment prior to TET oxidation; and the chemical-assisted pyridine borane sequencing (CAPS), in

which potassium perruthenate (KRuO4) is used instead of TET to only oxidise 5hmC, leaving all other cytosines unchanged (Liu et al. 2021).

Other emerging bisulfite-free methods include single-molecule real-time (SMRT) sequencing by Pacific Biosciences (PacBio), and nanopore sequencing by Oxford Nanopore Technologies (ONT), both offering direct readout of modifications and substantially longer reads (Flusberg et al. 2010; Levy and Myers 2016). These technologies are constantly being developed and already offer a powerful alternative to the aforementioned methods.

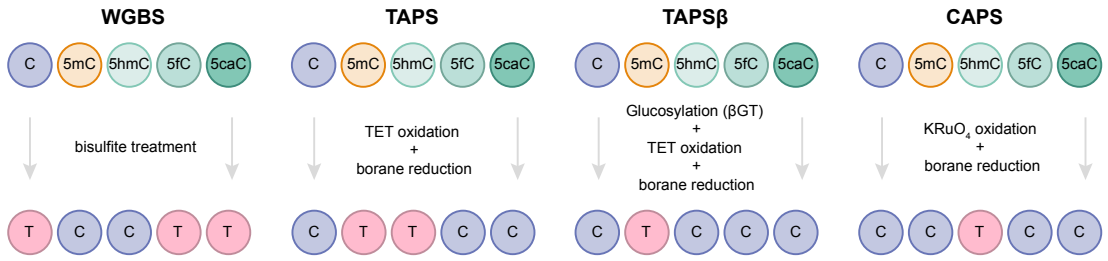


Figure 1.2: Summary of selected methods for base-resolution sequencing of cytosine modifications
Only the methods used to obtain data used in this thesis are included. WGBS converts unmodified and unprotected cytosines into thymines, while TAPS-based methods (TAPS, TAPSβ and CAPS) only convert modified cytosines of interest into thymines.

1.3 Genomic distribution and roles of 5mC

There are 32.8 million CpG pairs in the human genome (Gershman et al. 2022). Given the genome GC content of 0.45 (*Homo Sapiens Genome Assembly T2T-CHM13v2.0* 2023), it would be expected that CpG occurs with a frequency of $0.225 \times 0.225 = 0.05$, while the observed frequency is $0.0328 \div 3.17 = 0.01$. This phenomenon was observed long before the advent of whole genome sequencing and has been linked to increased mutagenecity of 5mC (Bird 1980). Specifically, methylated cytosines are more prone to be deaminated into thymines, resulting in a G:T mismatch (Lindahl and Nyberg 1974). If the mismatch is not repaired, it is fixed into C>T mutation during replication (Bellacosa and Drohat 2015). C>T mutations are the most prevalent mutations in cancer, somatic cells, and the germline (Alexandrov et al. 2013; Blokzijl et al. 2016; Rahbari et al. 2016; Ségurel et al. 2014).

CpGs tend to cluster in small regions named CpG islands (CGIs), which are typically associated with gene promoters and are predominantly unmethylated

(Bird 1984). The most commonly used criteria for classification of CGIs were proposed by Gardiner-Garden and Frommer 1987, and specify DNA regions with a high GC content (50% or higher), a length greater than 200bp, and a high ratio of observed CpGs to those expected based solely on the GC content (0.6 or higher). Slightly more stringent criteria were proposed by Takai and Jones 2002, and are used in some analyses. CGIs are 1000 bp - long on average, and are surrounded by "shores" and "shelves" of varying methylation states (within 2kb, or 2-4kb from the island borders, respectively) (Illingworth and Bird 2009). This leads to the phenomenon in which methylated CpGs are mostly found in the "open sea" or "CGI deserts," while unmethylated ones reside primarily in the CGIs. The distribution of the methylation states is bimodal, with most CpGs being either fully methylated ($\tilde{7}$0% of all) or fully unmethylated ($\tilde{3}$0%), and only a small fraction falling into the middle states. 5hmC has been found to be depleted in the islands and enriched in the remaining elements (Wilkins et al. 2019; Yu et al. 2012; Johnson et al. 2016).

The events of fertilisation and early embryonic development see two global demethylation and methylation events sweep across most of the genome (Guibert et al. 2012; Vincent et al. 2013). These dramatic changes ensure the correct cellular reprogramming and the establishment of appropriate imprinting and sex-specific methylation patterns (Greenberg and Bourc'his 2019). Following these events, the methylation patterns are established and remain stable, with the exception of local, cell-type-specific regulatory regions. The following section will aim to describe the 5mC landscapes at various genomic regions and summarise its roles in shaping gene expression profiles.

1.3.1 Promoters

Promoters are regions just upstream of genes, where relevant proteins, such as RNA polymerase (RNAP) and transcription factors (TFs), bind to initiate transcription. The level of promoter methylation is closely associated with its activity, although the mechanisms of these interactions are still unclear (Walsh and Bestor 1999). 5mC levels in the majority of promoters tend to be negatively correlated with the expression of downstream genes (Holliday and Pugh 1975; Wolffe and Matzke 1999; Jones and Takai 2001; Hsieh 1994). Approximately 70% of the promoters reside within CGIs, accounting for almost $\tilde{5}$0% of all CGIs (Saxonov et al. 2006; Illingworth et al. 2010). Very frequently, these promoters are linked to housekeeping genes, which are essential, highly conserved genes

ubiquitously expressed across tissues (Weber et al. 2007; Larsen et al. 1992; Zhu et al. 2008).

The CpG content of the promoters has a vital impact on its mode of interaction with the expression regulation system. CGI-associated, high-CpG promoters (HCPs) are characterised by very low methylation content, and while their methylation was found to repress certain linked genes (Velasco et al. 2010), they do not see an increase of methylation levels when inactive (Jones and Takai 2001; Weber et al. 2007; Bird 2002). Instead, they tend to acquire the H3K27me3 mark deposited by the Polycomb repressive complex 2 (PRC2), which is believed to be a dynamic control system allowing for a rapid induction or stable inactivation (Mikkelsen et al. 2007; Ku et al. 2008; Wu and Zhang 2015). Furthermore, unmethylated HCPs are often marked with H3K4me3 regardless of their activity (Wu and Zhang 2015). H3K4me3 was found to repel de novo methylation and is believed to be the cause of HCPs remaining in a hypomethylated state (Ooi et al. 2007; Thomson et al. 2010). In cancer, chromatin modifications have been found to precede CGI hypermethylation events (Strunnikova et al. 2005).

In contrast to HCPs, the expression of genes with low-CpG promoters (LCPs) (30% of all) has not been found to be correlated with 5mC levels (Weber et al. 2007). They tend to have high methylation levels and are associated with both active and inactive genes. LCPs are characterised by slightly different architecture; for example, they contain TATA boxes that are often lacking by HCPs (Reynolds et al. 1984) , and show more localised initiation of transcription, in contrast to a fairly dispersed one present in HCPs (reviewed in Juven-Gershon et al. 2008).

5hmC is strongly depleted in CGI and is therefore rarely found in the majority of promoters (Wilkins et al. 2019; Yu et al. 2012; Johnson et al. 2016). However, a negative correlation with gene expression has been found nevertheless (Pastor et al. 2011; Wu et al. 2011a).

1.3.2 Enhancers

Regulation of gene expression is not limited to their promoters, but also includes distal regulatory regions such as enhancers. First defined as DNA sequences that have the potential to increase basal transcription levels of genes at distances ranging from hundreds of bases to megabases, their mode of action remains elusive, but they are believed to enhance or enable transcription by recruiting transcription factors (TFs), RNA Polymerase II (RNAPII), and chromatin regulators

(Banerji et al. 1981; Lettice et al. 2003; Visel et al. 2009a; Chan and La Thangue 2001). Binding of TFs to their binding sites (TFBS) in enhancers encourages local nucleosome modeling and recruitment of many factors such as histone acetyltransferase p300 (HAT) (Chan and La Thangue 2001), or RNAPII followed by transcription of enhancer RNAs, whose role is still largerly unknown, but associated with the activity of given enhancers (Kim et al. 2010; De Santa et al. 2010). While promoters can be easily linked to the genes they regulate due to their close proximity, the distal nature of enhancer regulation provides challenges in determining the activity levels of enhancers and the associated genes. This is assuming we know the exact location of the enhancer in question. Multiple experimental and computational efforts were made to identify enhancers, resulting in a reliable signature of histone modifications and associations of other factors, including H3K4me1, H3K27ac and HAT (Visel et al. 2009b; ENCODE Project Consortium et al. 2007; Heintzman et al. 2007; Ernst and Kellis 2012).

Enhancers tend to reside in CpG-poor regions and are characterised by variable methylation levels (Kundaje et al. 2015). Active, TF-bound enhancers often display medium to low 5mC levels (Stadler et al. 2011; Ziller et al. 2013). The exact functional relevance of low 5mC levels at enhancers is unclear, and current research is facing two opposing theories. One suggests the instructive role of 5mC, where methylation actively decreases the affinity of TFs to bind to DNA (Tate and Bird 1993; Choy et al. 2010) or where methyl-CpG-binding domain (MBD) proteins bind to TFBS, blocking access to TFs. This model is not widely supported by experimental evidence. In contrast, experiments have found TFs which preferentially bind 5mC TFBS, leading to their demethylation (Stadler et al. 2011; Kress et al. 2006; Han et al. 2001). This case argues for methylation changes occurring downstream from TF binding. It is possible that a range of different TF-TFBS interactions happen, with 5mC having both an instructive and a passive role. Enhancers tend to be more cell-type specific than promoters (Ernst et al. 2011; Ernst and Kellis 2015; Kundaje et al. 2015), and lineage-specific enhancer methylation has been detected in T-cells (Schmidl et al. 2009). High levels of cell-type-specific expression is observed mostly in genes located in enhancer-rich regions of the genome, called super-enhancers (Whyte et al. 2013; Hnisz et al. 2013).

Tet proteins have been detected in active enhancers (Williams et al. 2011; Ficz et al. 2011), and these discoveries were followed by the identification of 5hmC enrichment at enhancer sites (Stroud et al. 2011; Johnson et al. 2016; Cui et al.

2020). Interestingly, they were found to be particularly enriched at very cell-type-specific enhancers. This suggests that 5hmC exists at enhancers in a transient state during the process of demethylation, further suggesting the noninstructive role of methylation. Conversily, the loss of Tet proteins leads to accumulation of 5mC in enhancers and their reduced activity (Lu et al. 2014; Hon et al. 2014).

1.3.3 Gene bodies

Having described the methylation states at promoters and enhancers, one cannot omit the introduction of the parts of the genome actually regulated by these elements: gene bodies. A pattern may be seen emerging from the previous sections, suggesting that 5mC plays different roles in different genomic contexts. The same stands for gene bodies, which are composed of exons and introns. First described in *Arabidopsis thaliana*, gene body methylation was found to be positively associated with transcription levels (Tran et al. 2005; Zilberman et al. 2007). This correlation was also confirmed to stand in humans, and subsequent studies suggested 5mC to be involved in the control of alternative splicing, spurious transcription and histone modifications (Maunakea et al. 2010; Rauch et al. 2009; Meissner et al. 2008; Faustino and Cooper 2003; Neri et al. 2017; Jeziorska et al. 2017). Due to the proximity to the promoter regions, the first exon and intron are usually excluded from the definition of the gene body. It was noted that similarly to promoters, their methylation is also associated with gene silencing (Brenet et al. 2011; Anastasiadi et al. 2018).

Approximately 1/3 of the CGIs in the genome are in gene bodies, of which 30–40% are methylated (Maunakea et al. 2010). Interestingly, methylated intragenic CGIs do not stop transcription (Jones 1999; Maunakea et al. 2010), which led to theories that 5mC is preferentially deposited in actively transcribed genes. Baubec et al. 2015 found that DNMT3B preferentially binds to bodies of transcribed genes supporting this theory. Gene body methylation was also proposed to be involved in the regulation of alternative promoters, under the assumption that most genes have two transcription start sites (TSS) (Maunakea et al. 2010). Methylation's role in splicing was suggested after observing that exons tend to have higher levels of modification to introns and the transitions occur at exon–intron boundaries (Laurent et al. 2010; Lister et al. 2009). High levels of methylation in actively transcribed gene bodies have been found to be associated with H3K36me3 histone modification (Ball et al. 2009).

Similar effects have been observed in the case of 5hmC. Multiple studies found its enrichment in gene bodies together with a strong positive correlation with transcription levels in multiple tissues (Wilkins et al. 2019; Song et al. 2011; Johnson et al. 2016; Cui et al. 2020). Results presented by He et al. 2021 suggest that the correlation is stronger than that of 5mC. Gene body 5hmC levels are cell-type specific and were found to change over time depending on the developmental context. For example, in adult livers 5hmC tend to occupy genes involved in catabolic and metabolic processes, in contrast to genes involved in differentiation and development pathways in fetal livers (Ivanov et al. 2013). Interestingly, genes that escape X-chromosome inactivation display higher gene body 5hmC levels than inactive genes (He et al. 2021). The exact role of 5hmC and the dynamic of methylation and demethylation events is still unclear, but evidence suggests that the intricate balance between methylation states is crucial for expression regulation beyond gene promoters.

It is also important to remember the mutagenic effects of 5mC mentioned before. Methylation at exons is the major cause of C>T mutations which may lead to diseases and cancer (Rideout et al. 1990; Schmutte et al. 1996). Conversely, high levels 5hmC were found to be associated with increased C>G mutation rate (Supek et al. 2014).

1.4 DNA modifications in cancer

There are two types of aberrant DNA methylation patterns found in cancer compared to normal cells of the same tissue. The first of them is global hypomethylation, especially enriched within broad late-replicating Lamin-associated domains containing many repetitive sequences such as Alu and LINE-1 (Ehrlich 2009). These CpG-poor regions cover over half of the genome, and remain mostly methylated in normal cells. Their methylation is crucial for chromatin organisation, repression of transcription, and maintenance of nuclear structure (Sen et al. 2006; Shapiro and von Sternberg 2005; Bodega and Orlando 2014). Hypomethylation was detected to begin early in tumourigenesis, and has been associated with tumour progression in multiple cancer types (Jackson et al. 2004; Suzuki et al. 2006). Both the origin and function of global hypomethylation is subject to an ongoing debate. Firstly, there is no consensus on whether the demethylation is caused by an active or passive process (or, possibly, a mixture of both) (Ehrlich 2009). Secondly, its function remains elusive – it has been proposed that hypomethylation at repeat regions may foster DNA rearrangements

that promote tumourigenesis, or loss of gene imprinting resulting in overexpression of associated, tumourigenic genes (Eden et al. 2003; Gaudet et al. 2003; Holm et al. 2005; Plass and Soloway 2002; Klutstein et al. 2016).

Studies of global hypomethylation are overshadowed by the second type of aberrant methylation patterns, that is of localised hypermethylation, especially at normally unmethylated CGIs. Affected CGIs are often located at gene promoters and may lead to silencing of tumour suppressor, or differentiation-associated genes (Esteller 2007; Widschwendter et al. 2007). Multiple established cancer genes were found to have hypermethylated CGI promoters leading to their silencing, such *VHL*, and *CDKN2A* (Herman et al. 1994; Cancer Genome Atlas Research Network 2012), and more recently studies expanded to whole genome analyses that identified whole pathways controlled by epigenetic aberrations (Jiao et al. 2014). The degree of CGI hypermethylation events varies between different cancers and even within cancers of the same organs (Saghafinia et al. 2018). Tumours with high frequency of CGI hypermethylation gained their own descriptive category called the CpG island methylator phenotype (CIMP) (Issa 2004). The extent and localisation of 5mC changes in cancer genome has been used in biomarker discovery and therapy guidance (Swisher et al. 2017; Locke et al. 2019).

Emerging evidence suggests aberrant profiles of 5hmC in multiple cancer types. Specifically, tumours present global decrease of 5hmC, often associated with the downregulation of *TET* mRNA transcription (Lian et al. 2012; Jin et al. 2011a; Kudo et al. 2012; Yang et al. 2013). In addition to global changes, studies in pancreatic cancer and glioblastoma found 5hmC to be enriched near regulatory elements such as enhancers and promoters and were associated with the upregulation of tumour-promoting genes (Johnson et al. 2016; Bhattacharyya et al. 2017).

1.5 Cell type-specific methylation and its applications

The intricate patterns of 5mC and 5hmC distribution across the genome and their evident association with the regulation of gene expression led to the investigations of differences in modification landscapes across different cells and tissue types. As indicated at the beginning of this chapter, cytosine modifications as well as the higher-order structures such as histones and chromatin, play a substantial role in the plastic definition of cell phenotypes. Despite most of genomic CpG positions

have the same 5mC status, methylation of some CGIs, gene bodies and enhancers varies depending on the cell's function and developmental origin. Methylation at different genomic contexts have varying levels of tissue-specificity. For example, CGIs at promoters rarely show unique methylation patterns due to their substantial association with housekeeping genes ubiquitously expressed across a variety of cell types (Song et al. 2005; Saxonov et al. 2006; Eckhardt et al. 2006). On the other hand, CGIs at gene bodies and intragenic regions have been found to display a greater tissue specificity (Maunakea et al. 2010; Illingworth and Bird 2009), . The majority of distinct methylation patterns has been found outside of CGIs, in the nearby regions called CGI shores - the tissue-specific methylation was found to be highly conserved and correlated with reduced gene expression (Irizarry et al. 2009). A high degree of tissue-specificity was observed in distal regulatory regions such as enhancers, which tend to be more cell-type-specific than promoters which methylation status is shared across cell types (Ernst et al. 2011; Ernst and Kellis 2015). Recent studies have indicated an even higher tissue-specificity of 5hmC, usually falling within introns and enhancers binding tissue-specific TFs (He et al. 2021; Nestor et al. 2012). The unique methylation patterns tend to be coupled with variable histone modifications, such as H3K4me1. Using ChromHMM, a method developed to to assign combinations of histone marks to unique genomic regions, the Roadmap Epigenomics consortium found the most tissue-specific marks to be present at regions surrounding TSSs, and in enhancers (Kundaje et al. 2015; Ernst and Kellis 2012).

Harnessing the methylation differences across cell types can be used in a range of applications, such as deconvolution to infer cellular compositions of a mixed sample. In such studies, sample methylomes can be deconvoluted using a range of refrence-free and reference-based methods to determine underlying similarities and cell proportions. In general, reference-free methods are used when there is no information available about the 5mC states of the cells contributing to the mixture, resulting in estimation of putative cell proportions which have to be assigned to correct populations using functional analysis. One of such methods was developed by Houseman et al. 2016 to study immune cell mixtures, using a modified version of non-negative matrix factorisation. Reference-based methods require the identification of differentially methylated regions (DMRs) specific to the cell types of interest, followed by the calculation of the contributions in studied mixtures using methods such as constrained projection/quadratic programming (Titus et al. 2017). These methods have been firstly developed to study immune cell proportions across a range of disease states, giving rise to the

field of *Immunomethylomics* (Houseman et al. 2016; Wiencke et al. 2017), and to account for confounding signals in epigenome-wide association studies (Jaffe and Irizarry 2014).

Cell type deconvolution has also found its application in the context of cancer, where reference-based methods were used to estimate tumour biopsy purity and levels of immune cell infiltration (Chakravarthy et al. 2018), and to find biomarkers associated with prognosis (Yang et al. 2017; Chen et al. 2015). Additionally, methylation markers have been crucial in the liquid biopsy research, where they can be used to assign the sources of otherwise nearly indistinguishable cell-free DNA fragments present in blood. For instance, Moss et al. 2018 developed a reference-based approach for the identification of cell-free DNA originating from a range of normal tissues and cancers, including those of unknown primary.

The referenced studies are primarily based on methylation arrays, which restrict the analysis to either 450,000 or, in more recent investigations, 850,000 CpG sites. Up until very recently, the availability of whole-genome methylation datasets was limited to public data obtained by the ENCODE and Roadmap Epigenomics consortia (The ENCODE Project Consortium 2012; Kundaje et al. 2015), which suffer from a lot of intra-sample heterogeneity, especially in case of bulk tissues (Loyfer et al. 2023). There was no other integrated whole-genome atlas of methylation representing a range of human tissues. This thesis aims to bridge this gap, by truly whole-genome analysis of unique methylation patterns, and its potential application in deconvolution of cancer data.

1.6 Aims of the thesis

1. Utilise the availability of TAPSβ and CAPS whole-genome maps obtained from multiple human tissues to create a new reference for tissue-specific methylation patterns, in a semi-supervised way.
2. Investigate the tissue-specific methylation patterns for the evidence of biological relevance and 5hmC contributions.
3. Deconvolute oesophageal tumour biopsies using the constructed reference atlas and analyse the biomarker potential of differences in cellular contributions.

2

Method development

Contents

To develop a method to obtain tissue-specific methylation patterns, I used a TAPSβ atlas representing a collection of 20 different cells and tissues. This dataset provides a comprehensive single-base resolution map of 5mC, and together with the complementary CAPS data from the same samples, significantly enriches our understanding of methylation patterns. This chapter explains the development of this reproducible pipeline, which is used to obtain results analysed and used in the following chapters.

2.1 TAPSβ atlas

2.1.1 Data preparation

The original TAPSβ atlas consisted of 116 samples representing 34 healthy tissues, blood cell populations, several tumours, and pancreatitis. Sample preparation, sequencing, and initial processing were performed by colleagues, and are summarised in Methods 6.1. I focused on non-disease tissues, and removed samples with abnormally low coverage, quality metrics or potentially mislabeled samples (data not shown). The remaining atlas consists of a set of 71 samples representing 20 tissues and cell types (Figure 2.1), which are summarised in the Appendix A.1.

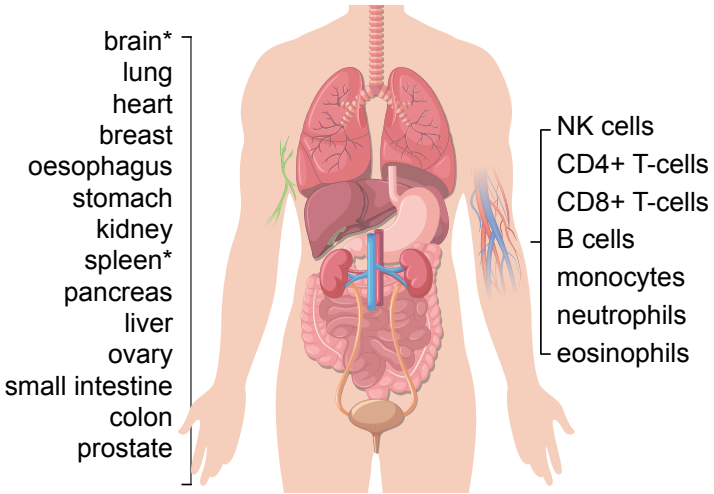


Figure 2.1: TAPS atlas and pipeline summary
The various tissues and cell types incorporated in the TAPSβ atlas. Some samples from tissues marked with an asterisk (*) were removed in later processes because of substantial discrepancies in their methylation landscapes.

To reduce the size of the datasets, I merged CpG pairs from the forward and reverse strands, under the assumption that most of CpG methylation is symmetrical. I removed CpGs in regions considered as "problematic" by the ENCODE consortium, which cover sequences that tend to have anomalous, unstructured, or high signal in sequencing experiments (Amemiya et al. 2019). Additionally, I removed any sites in centromeres and sites covered by less than 5 reads in each sample. Only autosomal chromosomes were included in the analysis.

Next, I removed CpGs that are likely to be single-nucleotide variants (SNVs). During the process of TAPSβ sequencing, 5mC is converted to T. The abundance of C>T SNVs in the genome may influence the methylation calling results, reporting

that CpGs are methylated, while they are, in fact, TpGs. This can be addressed by looking at the base opposite to the C in a CpG, which should be a G in the case of cytosine modification and A or C in the case of C>T or C>G substitutions, respectively. I calculated the proportion of Gs on the opposite side of Cs in CpGs applying a modified version of the alleleCount function on BAM files from each sample (described in detail in Methods 6.1.1) (*alleleCount* 2023). If the proportion was lower than 0.6, I assumed that this is a heterozygous or homozygous SNV. On average, the samples had 463,915 SNVs each (not accounting for differences in sequencing coverage) (Supplementary Figure B.1). Of these, on average, only 11.13% were covered in the dbSNP database v154 (Kitts and Sherry 2011). This supports the use of this method instead of just filtering out common SNPs, as it removes a large proportion of potentially confounding sites. The affected CpG sites were removed from the corresponding methylation maps.

2.1.2 Genome segmentation

To identify tissue-specific sets of CpGs, I first divided the genome into a set of continuous non-overlapping blocks with consistent methylation states in each tissue. This was done to minimise the effects of methylation states of single CpGs, and to increase the power of the analysis by adding the modification states across the segments, which can be necessary in, for example, cell-free DNA analyses. Additionally, it compresses the data for computation. The segmentation was done under the assumption that 5mC's role in genome regulation is via sets of co-operating, neighbouring sites, rather than individually.

Firstly, I merged samples from the same tissue or cell type. For each CpG, I combined modified and unmodified cytosines and calculated the average beta value per tissue group. The tissue assignment is summarised in Table A.1. To allow for small variation of methylation of CpGs across tissues, each site was assigned to one of three groups, depending on its beta value: 0-0.25 as "unmethylated", 0.25-0.75 as "hemimethylated", and 0.75-1 as "methylated". I merged the data for each tissue group into one table, removing CpGs missing in more than 25% of the groups.

To identify segments of uniformly methylated sites, I first divided this table by chromosomes, and further into 10-kb long segments to ensure comparable clustering and to speed up the computation. On each segment, I performed hierarchical clustering to identify sets of sites with identical methylation states, as illustrated in Figure 2.2. The dendrogram was cut at the height of 0.5, assigning

CpGs of the same methylation patterns across all tissues to the same clusters, with a small degree of flexibility that allowed handling of the sporadic missing values. Then, under several restrictions listed below, I identified blocks of consecutive CpGs that belong to the same clusters. The distance between consecutive CpGs could not be greater than 1 kb and so was the total span of the block. Additionally, I required at least three CpG sites per block to ensure adequate coverage of the genome while preserving the comethylation of the segments. Because of the removal of some CpGs prior to block selection (due to low coverage or SNVs), some blocks could contain more CpGs than identified. To address this, I compared the observed and expected number of CpGs in each block using a whole genome CpG map and excluded all blocks with more than 30% CpGs which were unaccounted for. The segmentation of the genome according to these requirements yielded 884,842 blocks, covering 5,741,031 individual CpGs. The distribution of the blocks is highly skewed toward the shortest blocks, as they make up 42% of the identified segments.

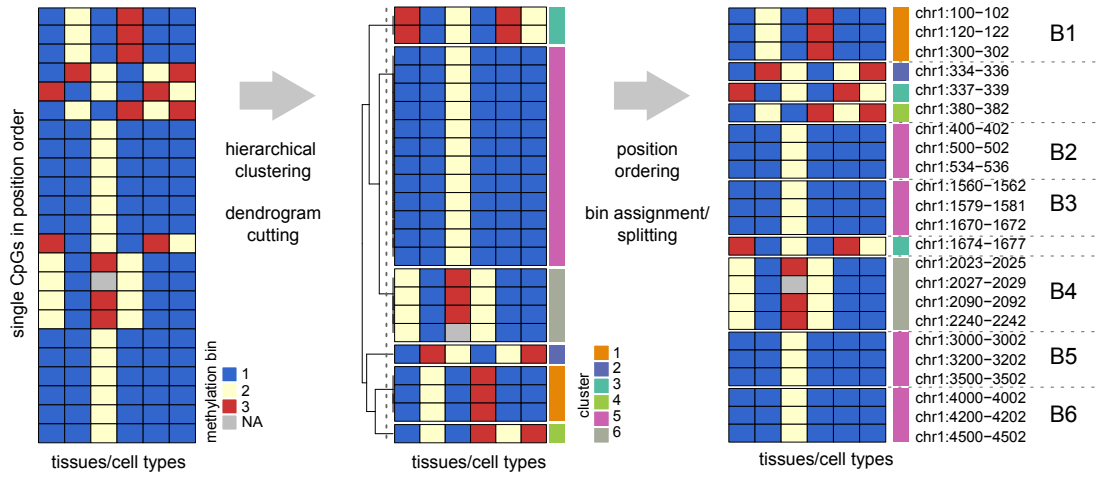


Figure 2.2: Summary of the genome segmentation process
A representative illustration of the process of genome segmentation. First, the samples are merged based on the tissue of origin and beta values are converted to methylation level bins (not shown). Then, 10kb-long sections are clustered and the dendrogram is cut at 0.5 height (dotted vertical line), assigning each CpG to a cluster. CpGs are then reordered back according to the location in the genome, and neighbouring CpGs from the same cluster are assigned to one block. The figure shows the following scenarios: B1, a canonical block; B2&B3, blocks separated because of the >1kb distance between 3th and 4th CpG; B4, a block with a missing value; B5&B6, block split in two because of the length >1kb.

2.1.3 Variance filtering

Having identified the blocks co-methylated CpGs, I sought to eliminate regions sharing the same methylation values in all tissues. For this and all subsequent steps, I considered all samples individually, without grouping them by tissue/cell type. Using the positions of all identified blocks, I summed modified and unmodified values of the CpGs across each block in each sample and calculated the beta values. If a given region was not covered in a particular sample, I imputed the beta value based on the median value of all samples for that block. This approach assumes that the missing value will adopt a "neutral" methylation state, which is the most common methylation state in all tissues. Of 884,842 total blocks, each tissue was missing on average 3,050 blocks. 45% of the missing blocks were missing in one sample only.

Because of the nature of methylation, most of these blocks have very similar methylation states in all samples. To eliminate these noninformative sites, I filtered sites based on the per-block variance across tissues. I tested several variance cutoff values, and guided by the obtained numbers of blocks, previous papers, and downstream analyses, I chose 0.01 as the optimal cutoff (Table 2.1). Due to the substantial difference in methylation profiles between tissue- and blood-derived samples (as illustrated in the heatmaps below), I noted that variance filtering separately in both groups results in a higher number of identified blocks (Figure 2.3). As a result, I obtained 58,004 high-variance blocks.

Variance filter	Number of identified blocks
>0.05	3,669
>0.01	58,004
>0.005	116,785
>0.001	762,924

Table 2.1: Number of blocks obtained with different variance filters

The methylation landscape of the selected blocks is illustrated in Figure 2.4. This heatmap illustrates the different global methylation patterns in tissues and cells isolated from blood, excluding samples originating from brain. The majority of the identified blocks are hypermethylated across most samples and hypomethylated in small subsets, suggesting that hypomethylation plays a more significant role than hypermethylation in portraying cell type specificity. Interestingly, when I ran the entire process including brain samples, I got a substantially different methylation landscape, and a considerably lower number of high-variance blocks

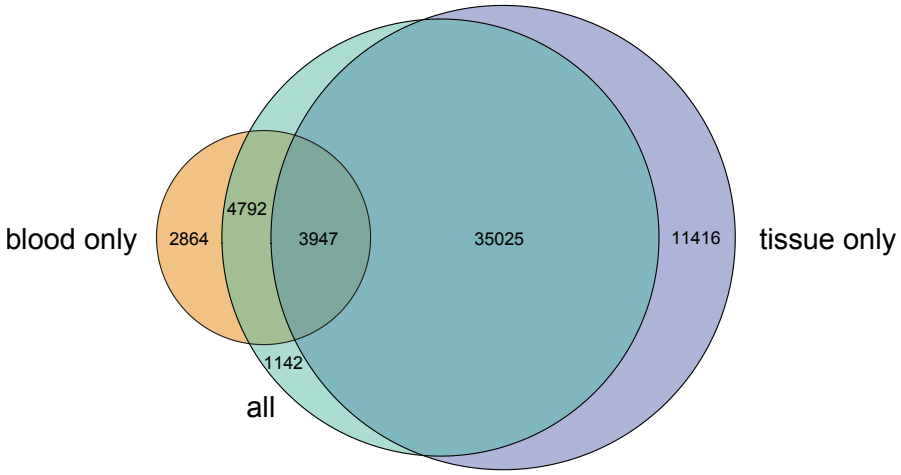


Figure 2.3: Increased block numbers via separate filtering of blood and tissue samples
Separate variance filtering of blood and tissue samples (represented by green and orange circles, respectively) results in a substantially larger number of retained blocks compared to the combined filtering approach (depicted by the blue circle).

(41,073, in contrast to 58,004) (Figure 2.5). Brain tissue exhibits a distinctly low methylation profile, representing a considerably higher proportion of sites compared to other tissues. Taking into account the distinctiveness of the brain methylation landscape and the well-established substantial contribution of 5hmC to brain tissue (Kriaucionis and Heintz 2009), I decided to not include the brain samples in further analyses.

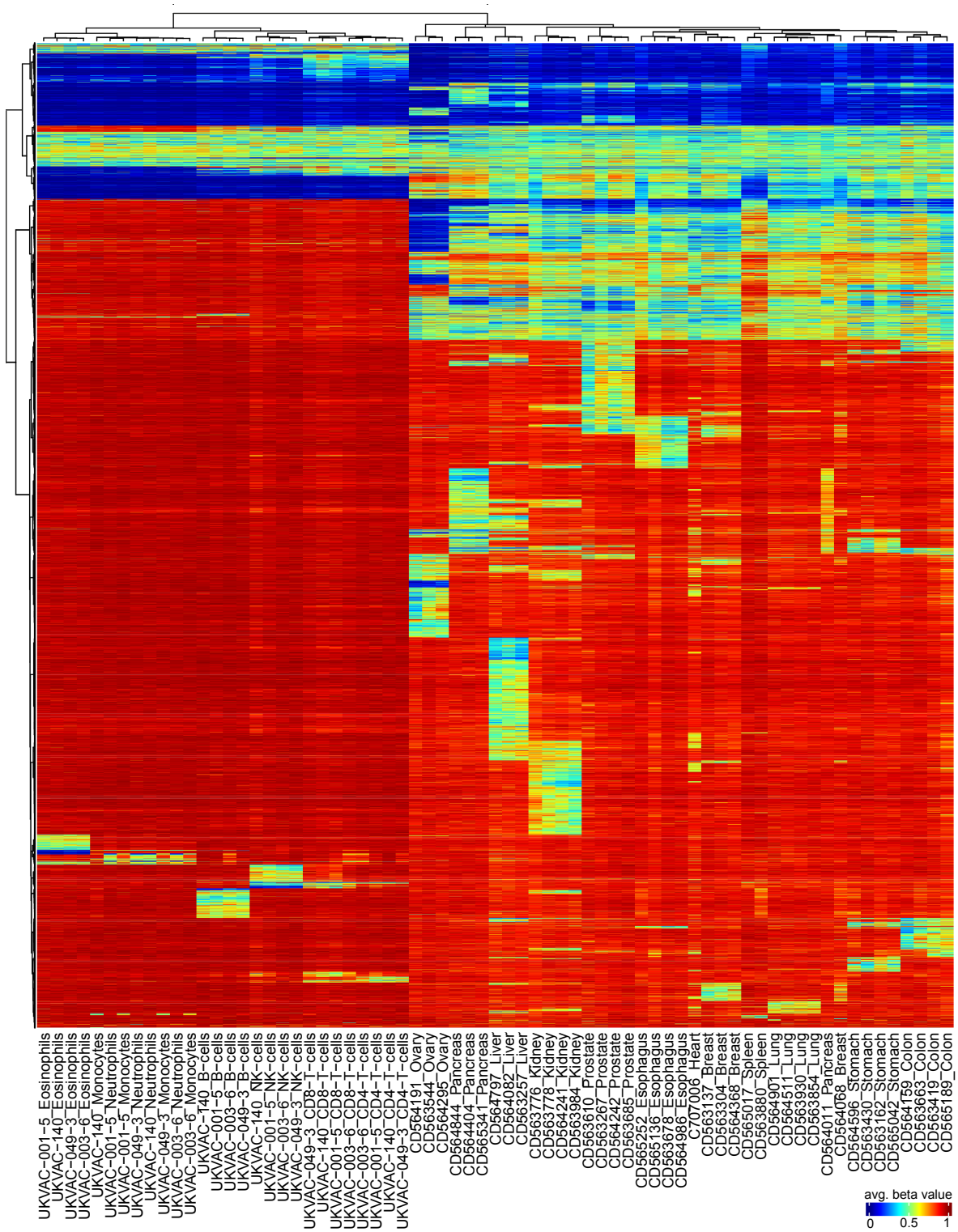


Figure 2.4: Methylation status of highly-variable blocks in the TAPSβ atlas
Beta values of blocks with a variance greater than 0.01 per block, computed separately for blood and tissue samples, resulting in a total of 58,004 highly variable blocks. The data were clustered using the Ward's method.

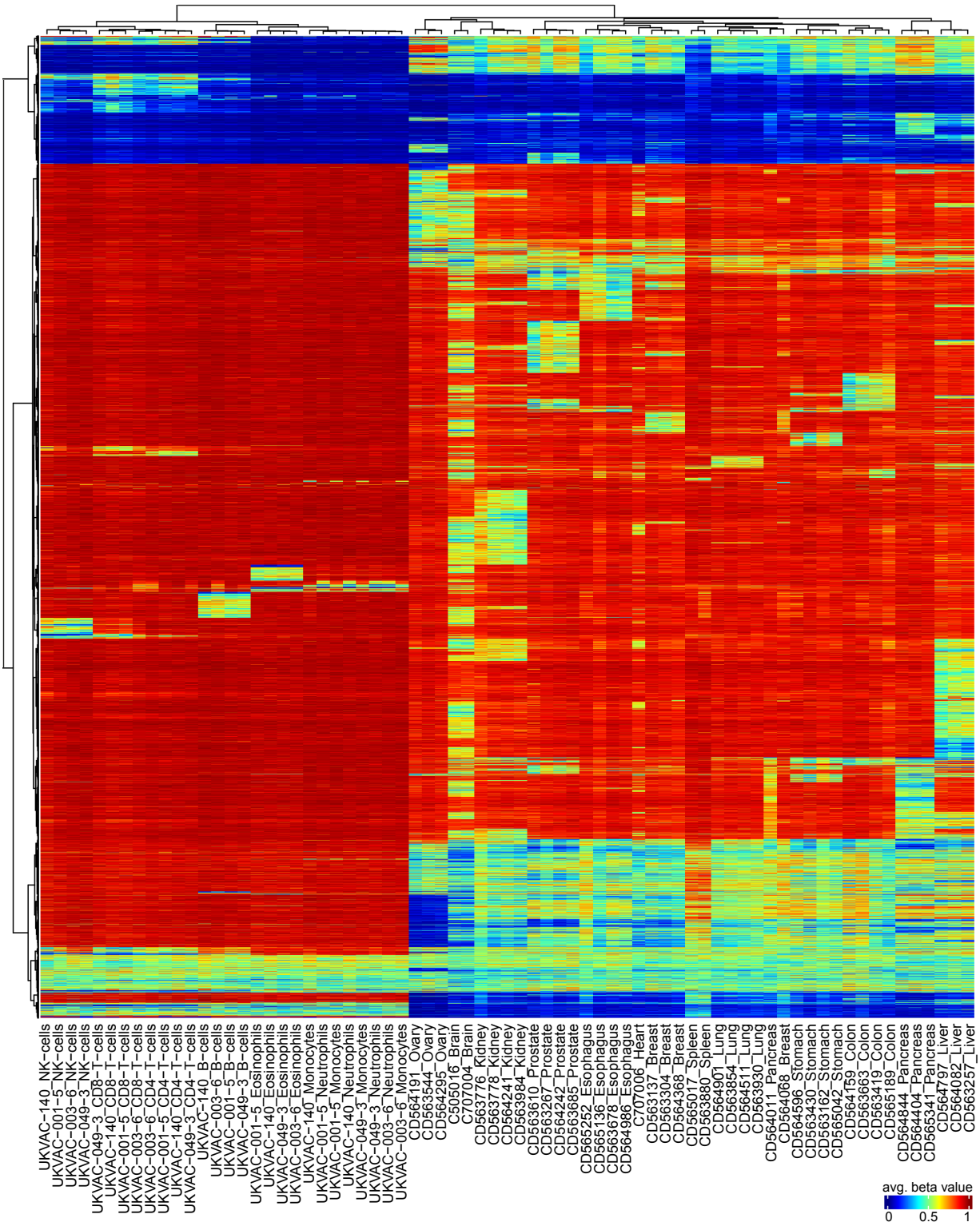


Figure 2.5: Methylation status of highly-variable blocks in the TAPSβ atlas, including brain samples

Beta values of blocks with a variance greater than 0.01 per block, computed separately for blood and tissue samples, resulting in a total of 41,073 highly variable blocks. A unique methylation pattern can be seen in brain samples. The data were clustered using the Ward's method.

2.1.4 Properties of the selected CpG blocks

The selected properties of the CpG blocks are detailed in Table 2.2. Compared to all the blocks identified during the genome segmentation process, I found that highly variable blocks were typically smaller and contained fewer CpGs. Moreover, these blocks exhibited a higher average distance from each other, and fewer blocks were found within 100bp from each other. This indicates a pattern of less clustering among the highly variable blocks compared to all blocks identified across the genome.

	All Blocks	High-variance
Total Number of Blocks	884,842	58,004
Number of Covered CpGs	5,741,031	211,572
Median Number of CpGs per Block	4	3
Mean Number of CpGs per Block	6.42	3.57
Median Length of Blocks	116	51
Mean Length of Blocks	213.7	85.42
Median Distance Between Block, bp	1,182	15,019
Mean Distance Between Blocks	6,013	47,162
Blocks Less Than 100bp Apart	106,224	2,865
Blocks Less Than 100bp Apart, %	12%	4%
Average Blocks with Missing Values, Per Sample	4,162	394
Average Blocks with Missing Values, Per Sample, %	0.4%	0.67%

Table 2.2: Genome segmentation and variance filtering summary
The values represent absolute numbers, unless otherwise noted. CpGs refer to pairs of C-Gs on both strands of the DNA. High-variance blocks refer to the dataset without brain samples.

To better understand the landscape of the regions represented by the high-variance CpG blocks, I investigated the degree of overlap between genic regions, CGIs and enhancers, and CpGs representing the whole genome, all identified blocks and highly variable blocks (Figure 2.6 A-C). Assignment into the three categories was performed independently, meaning that, for example, a given CpG can be in both a CGI and a promoter. Nearly all CpGs overlapped at least one region. The majority of CpGs in all groups fell into intergenic, intronic, or open sea regions, as expected by their general size in the genome (represented by the orange bar). Out of all high-variance blocks, only 6356 (containing 22,541 CpGs) were in the open sea CpG and did not overlap with any genic or enhancer region. Detailed methods of the analysis are described in Methods 6.3.

My approach of selecting co-methylated groups of CpGs may result in an over-representation of certain genomic regions, such as CGIs, because they are CpG-dense and usually unmethylated. This can be clearly seen from the difference in proportions in Figure 2.6 A-C). To fully quantify the over-representation at each region on each filtering stage, I compared the observed and expected proportions of covered regions using Fisher's exact test (Figure 2.6 D-F). CpG islands are clearly preferred during the genome segmentation stage due to their homogeneity, but are strongly selected against during variance filtering (OR 18 vs OR 0.01, p val. <0.001). Compared to the whole genome, high-variance blocks are preferentially located in CpG shores and shelves, rather than islands and open-sea regions. The patterns in genic regions strongly resemble those of the CGIs, with high selection for promoters and 5' UTRs at the stage of segmentation, which is lost during variance filtering. Highly variable blocks are significantly enriched for introns, exons, 3' UTRs and regions 1 to 5kb upstream of TSS (OR > 1.1, p val. <0.001). There is a strong selection against enhancers during genome segmentation, which is then overcome during variance filtering. The above results suggest that, despite strong selection for certain regions during the genome segmentation process, the variance-filtering step recovers sites which are more expected to be variable between tissues. Interestingly, only 4% of the high-variance CpG blocks are included in the Illumina EPIC array, highlighting the benefits of using whole-genome data.

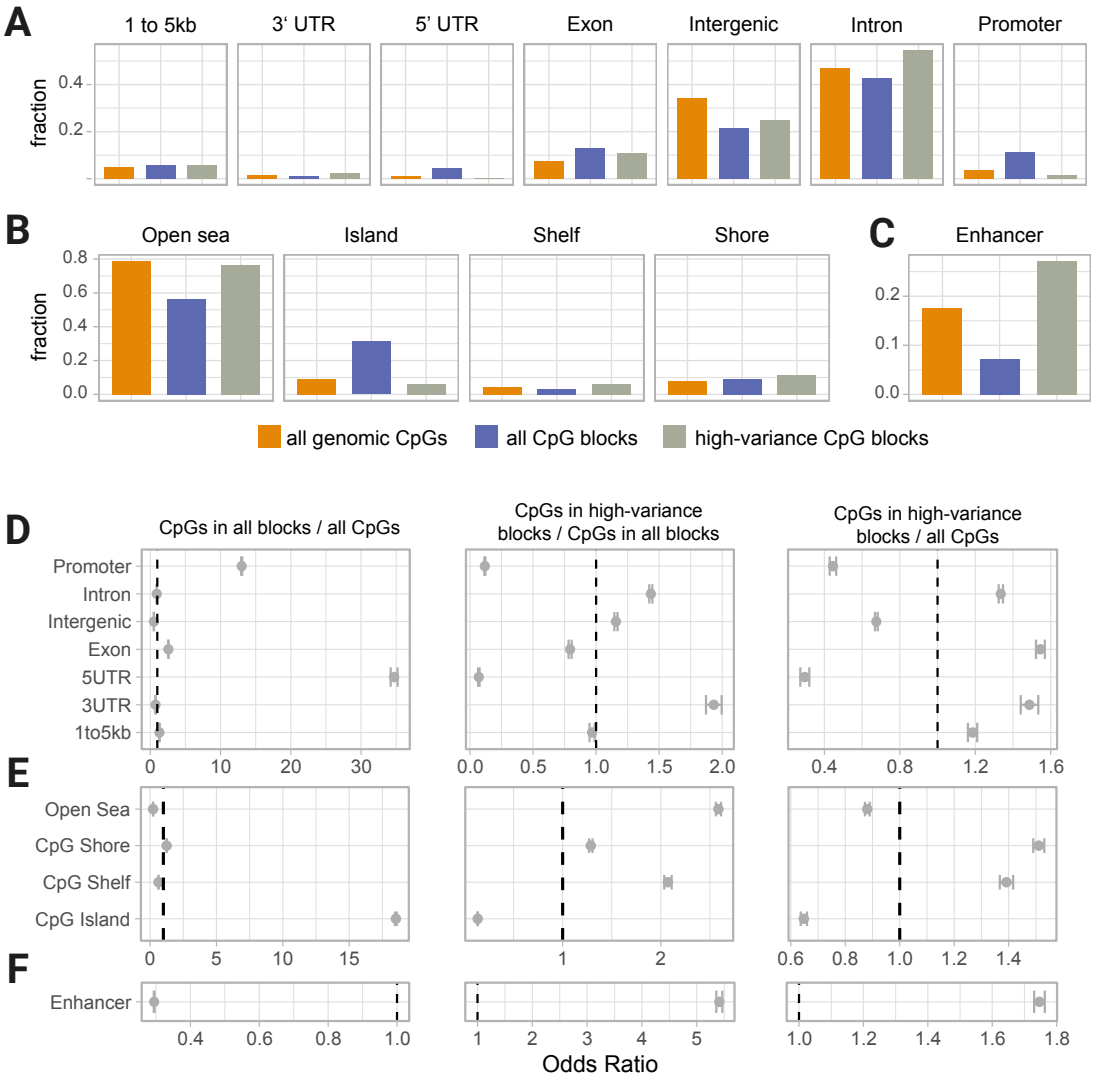


Figure 2.6: Representation of genomic regions during the segmentation and filtering processes.
The CpGs captured during the genome segmentation and filtering process represent different genomic regions, as illustrated in panels A-C. The bars represent the fractions of CpGs in each region. The analysis was performed separately for each panel to account for overlaps between regions, and the overlaps were calculated using individual CpGs. A. Genic elements. The majority of CpGs in blocks in cover intergenic and intron regions. Only transcripts included in the MANE database are included for clarity. 1 to 5kb, 1-5Kb upstream of the TSS. B. CGI elements. C. Enhancers. Only non-promoter enhancers were included. I used Fisher's exact test to assess the over-representation of genomic elements during block selection processes (D-F). A strong preference for homogeneously methylated regions such as CpG islands, and elements associated with them such as promoters and 5' UTRs is observed at the genome segmentation stage (column 1). This effect is reversed during the variance filtering stage (columns 2 and 3). The selected, high-variance CpG blocks are strongly enriched for CpG shores, shelves, introns, exons, 3' UTRs, regions 1 to 5kb upstream of TSS, and enhancers. The ratios were calculated using the total number of CpGs assigned to the blocks, within or outside a given genomic region. The error bars represent 95% confidence intervals, and the dashed line indicates a odds ratio of 1. All presented ratios had p values of <0.001.

2.2 Non-negative matrix factorisation

The identification of the highly variable blocks was the first step in obtaining a map of tissue-specific CpG blocks. Then, I had to identify and divide the blocks that are specific for each tissue in the atlas. To do this, I reduced the dimensionality of the complex data of the methylation values per CpG block per sample, using non-negative matrix factorisation (NMF).

NMF is one of the algorithms used for the factorisation of matrices representing complex multidimensional datasets, similar to principal component analysis (PCA). The objective of NMF is to explain the observed data using a limited number of basis components which, when combined together with the estimated coefficients, approximate the original data as accurately as possible.

To use NMF to identify tissue-specific methylation patterns from our set of samples, I generated a matrix A consisting of the methylation levels of p CpG blocks in n samples representing various tissue and cell types (simplified overview in Figure 2.7. The actual input matrix is presented in Figure 2.5). The goal is to find a number of k dimensions (or "signatures"), ideally corresponding to the number of tissue groups in the input, each defined as a positive linear combination of the p features. Then, the methylation patterns of the samples can be approximated as positive linear combinations of these dimensions.

This corresponds to factoring matrix A into two matrices such that:

$$A \approx WH$$

The matrix W has size pk, with each of k columns defining a methylation signature, and each entry W_{ij} being the coefficient of CpG block i in signature *j*. The coefficients can be thought of as "weights" of each feature for each signature. Matrix H is of size nk, with each of the n columns representing the signature contribution pattern of the corresponding sample. The entry H_{ij} represents the exposure of the signature i in sample j (Figure 2.7).

The method randomly initialises the matrices W and H and iteratively updates them to minimise their divergence from the original matrix A. The algorithm may not converge to the same solution on each run. If the clustering in the k clusters is stable, the samples are assigned to the same clusters with each run. The stability of the clusters can be inspected by looking at the cophenetic correlation

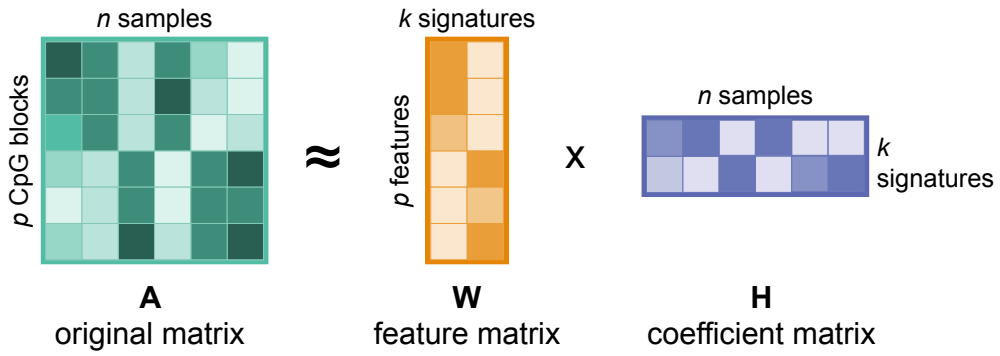


Figure 2.7: Graphical overview of NMF
Intensity of the colour corresponds to methylation values in matrix A, weight of coefficients in matrix W, and relative signature contribution in matrix H.

coefficient, which, in short, reflects the probability that samples will be assigned to the same clusters in each run (Brunet et al. 2004).

The use of NMF in the detection of tissue-specific methylation patterns from a set of samples differs from the original application to deconvolute mixtures of signals. Assuming that our input samples come from a single tissue or cell type, and I filter out nonvariable CpGs, our signals are already deconvoluted. For this purpose, the decomposition of matrix A can help us isolating truly tissue-specific CpGs in the form of matrix W, and the contribution of each signature to the samples in the matrix H. Ideally, I want k to be equal to the number of tissue types present in the data set, but this may not always form the most stable cluster with the highest cophenetic value. The biological relevance of the clusters and the stability have to be taken into account, because any sample heterogeneity or a higher level of heterogeneity in the data (such as the presence of generally heterogeneous tissues together with a more pure blood population) can lead to lower values of k being more stable.

Furthermore, I had to address the nature of the representation of methylation values, which is a proportion rather than the counts expected by NMF. If in a row p the majority of values are 1, with a few samples at 0, this block will not be considered "important" for any signature, due to the nature of the algorithm. However, this example represents the most common situation in our input matrix, in which a CpG block is uniquely hypomethylated in a given tissue. To mitigate this issue, which would inevitably dramatically reduce the number of important regions for each signature, I simply transformed the input matrix rows with mostly methylated blocks such as $a_ij = abs(1 - a_ij)$, to present them as a generally unmethylated, with hypermethylation at the particular tissues of interest.

2.2.1 NMF on the tissue atlas

Having identified the highly variable blocks, I used NMF to identify patterns in CpG blocks that are significant in characterising each tissue/cell type. Choosing a factorisation rank of 19, equivalent to the number of distinct groups in the atlas, allowed us to assign unique signatures to almost all samples of the corresponding tissues (Figure 2.8). This resulted in a near-perfect match for several tissues, including the oesophagus, stomach, colon, and other solid tissues.

Blood-derived samples, which had fewer high-variance input CpG blocks, showed less granularity in deconvolution. Neutrophils and monocytes were assigned the same signature (signature 14), which is not unexpected considering their shared developmental lineage. However, a more monocyte-specific "sub" signature (signature 5) was observed that further distinguishes them. Similarly, CD4+ and CD8+ T cells were not fully separated, with signature 9 being the primary signature. Furthermore, CD8+ T cells had an additional contribution from signature 6, interestingly shared with two of the natural killer (NK) cell samples.

There were no signatures specific to a subset of patients, rather than tissues, except a small trace of signature 14 in each sample from the UKVAC-003-06 donor. This suggests the successful removal of potential confounding SNPs, which improves confidence in the tissue specificity of the observed methylation signatures.

Choosing the appropriate factorisation rank for the analysis was crucial to achieving accurate results. This choice was largely guided by the number of tissue/cell types present in the atlas, and was further confirmed by the high stability of the clusters, as indicated by the quality statistics of the NMF (Figure 2.9). A factorisation rank of 19, corresponding to the number of distinct groups in the atlas, was observed to identify the most stable clusters.

Additionally, I explored how NMF behaves at higher and lower ranks to understand the underlying patterns of the data. When the rank was lower, it failed to clearly distinguish between certain cell types. Specifically, neutrophils and monocytes, as well as CD4+ and CD8+ T cells, were no longer separated (Figure 2.10 A). On the other hand, higher ranks led to a complete separation of some samples or the creation of additional sub-signatures (Figure 2.10 B).

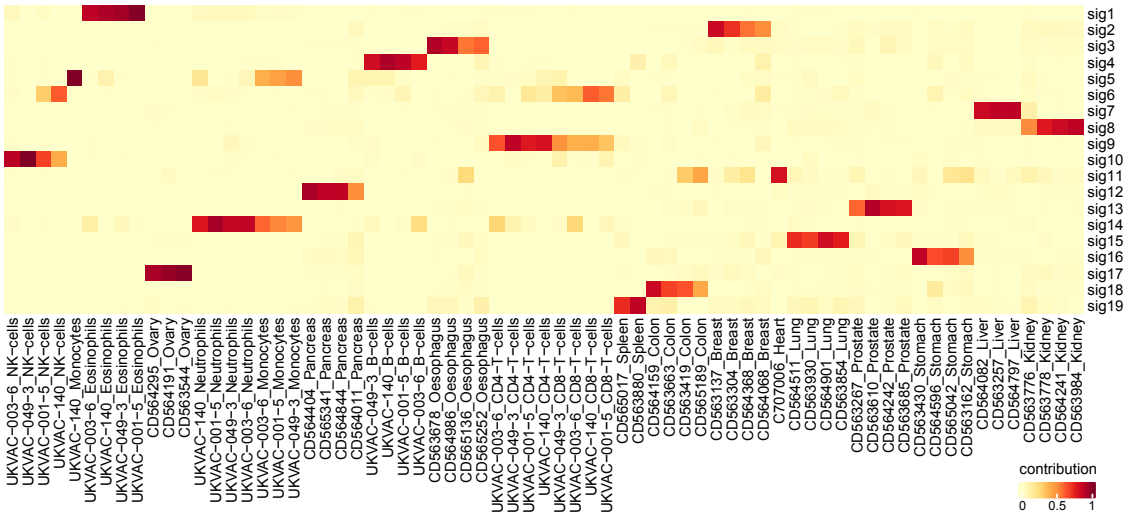


Figure 2.8: Strength of contribution of each identified methylation signature to each sample
Each methylation signature identified through NMF contributes in varying degrees to the overall methylation pattern of individual samples. Samples originating from the same tissue or cell type are mainly assigned the same signature, confirming the detection of tissue-specific methylation signals. Two pairs of cell types – Neutrophils/Monocytes and CD4/CD8 T-cells – share the same principal signature but also exhibit a secondary "sub-signature" that distinguishes the cells within each pair.

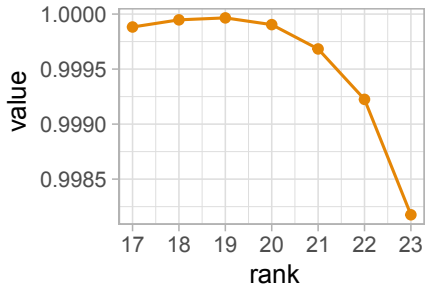


Figure 2.9: Cophenetic value of NMF runs at different factorisation ranks
The most stable clusters are at rank 19, which corresponds to the number of tissues included in the atlas.

2.2.2 NMF coefficient analysis

To understand which CpGs have the highest contribution to each signature, I investigated the coefficients (later referred to as "weights") of each block to each signature (Figure 2.11). Weight histograms show distinct shapes for signatures derived from tissue and blood samples. In the former, the weights tend to follow a bimodal distribution, with a pronounced minor mode composed of a large number of high-weight blocks. Blood signatures, on the other hand, show a less pronounced minor peak, and this appears to be composed of a smaller number of

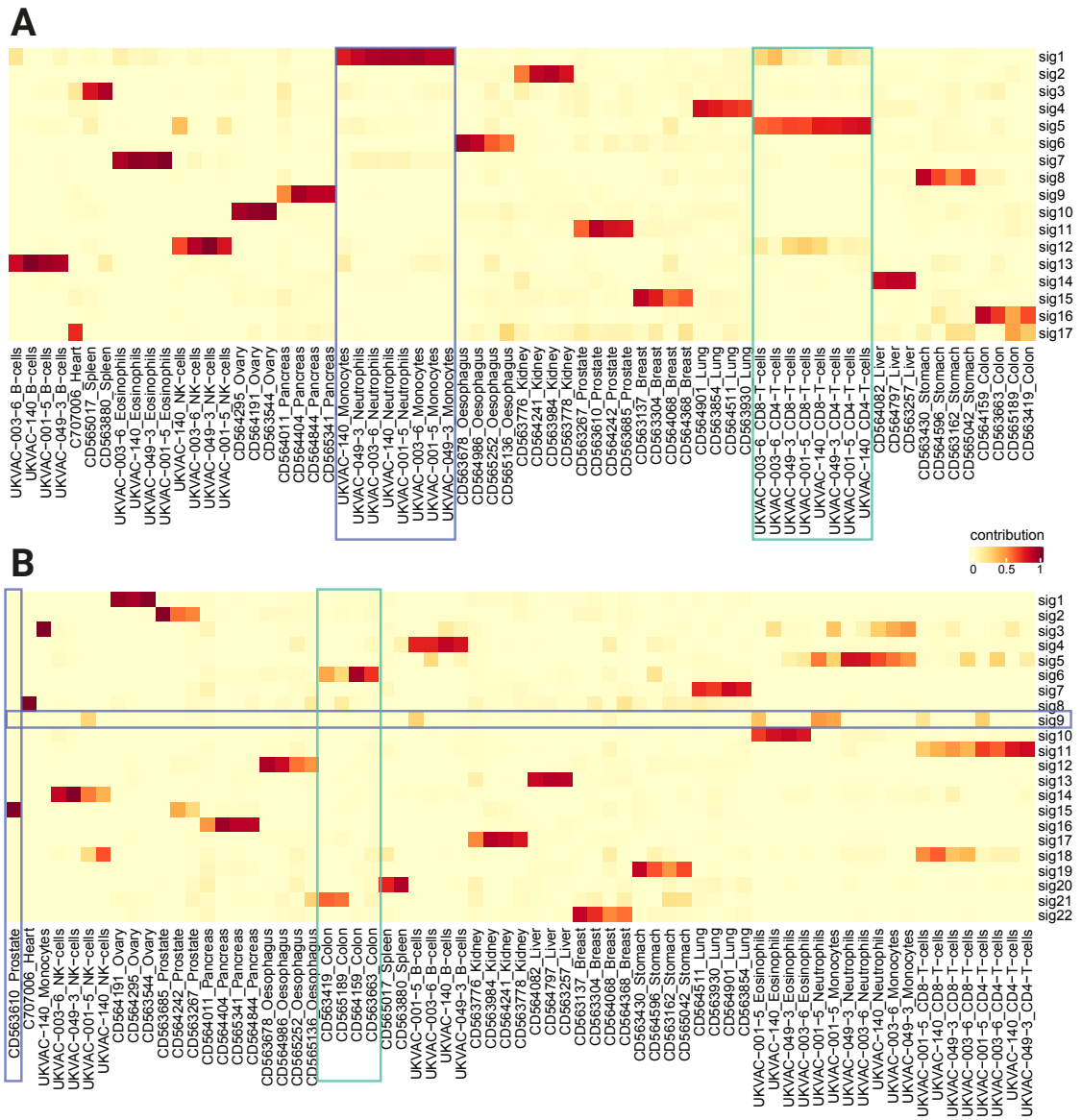


Figure 2.10: Alternative factorisation ranks fail to assign correct signatures
A. Contribution matrix illustrating the results of NMF runs at ranks lower than the number of tissues, failing to distinguish similar blood cell types. B. Contribution matrix showing the results of NMF runs at ranks higher than the number of tissues, revealing an isolated prostate sample and the creation of artefact signatures corresponding to a single patient (sig9) or isolating colon samples (sig21).

blocks of relatively higher weight compared to tissues. I applied the Expectation Maximisation (EM) algorithm to select the cutoff between the distributions, to identify high-weight blocks in each signature in an unbiased manner (Methods 6.2). From this point on, the terms "high-weight" or "important blocks" will refer to those sites that were above the cutoff point in each signature.

As illustrated in Figure 2.12 A, the number of high-weight blocks in blood signatures is considerably smaller than in tissue signatures. Interestingly, the spleen signature, which is expected to have a high blood content due to the nature of the organ, has fewer informative blocks compared to other tissues. Only a small fraction of the total informative blocks per tissue consist of hypermethylated sites, and contribute significantly more to blood signatures compared to tissue signatures (Wilcox rank sum test, p.val <0.001, Figure 2.12 B).

To understand the distinct difference in the numbers and weight distributions of the selected blocks between the two groups of signatures, I investigated the genomic locations of the selected CpGs in each group. As shown in Figure 2.12 C, a higher proportion of blocks in the blood group are located in CpG islands and shores, promoter-associated enhancers, promoters, 5'UTRs and regions 1 to 5kb upstream of TSS. These genomic annotations are typically associated with promoter regions. Conversely, there is a significantly smaller contribution of blocks overlapping with inter-CpG island regions, enhancers, introns, and 3' UTRs.

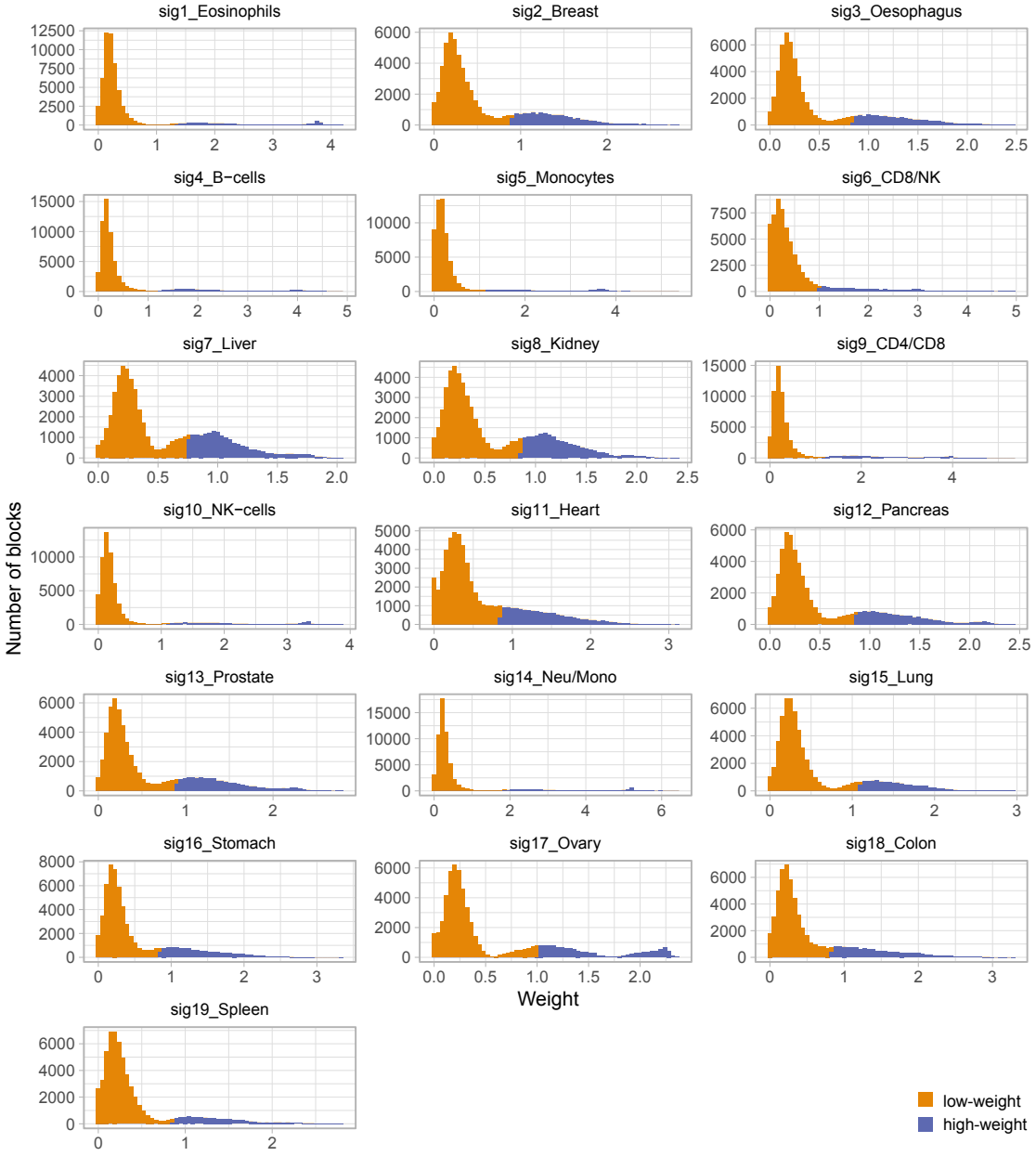


Figure 2.11: Weight distribution varies across signatures
Tissue-type signatures display a pronounced secondary peak that contains many high-weight CpGs, while blood types usually exhibit a lower number of high-weight blocks within a flatter, more dispersed secondary peak. The highlighted areas represent the blocks considered informative for each signature. Signature names were assigned based on tissues with the highest contribution to a given signature, as shown in Figure 2.8.

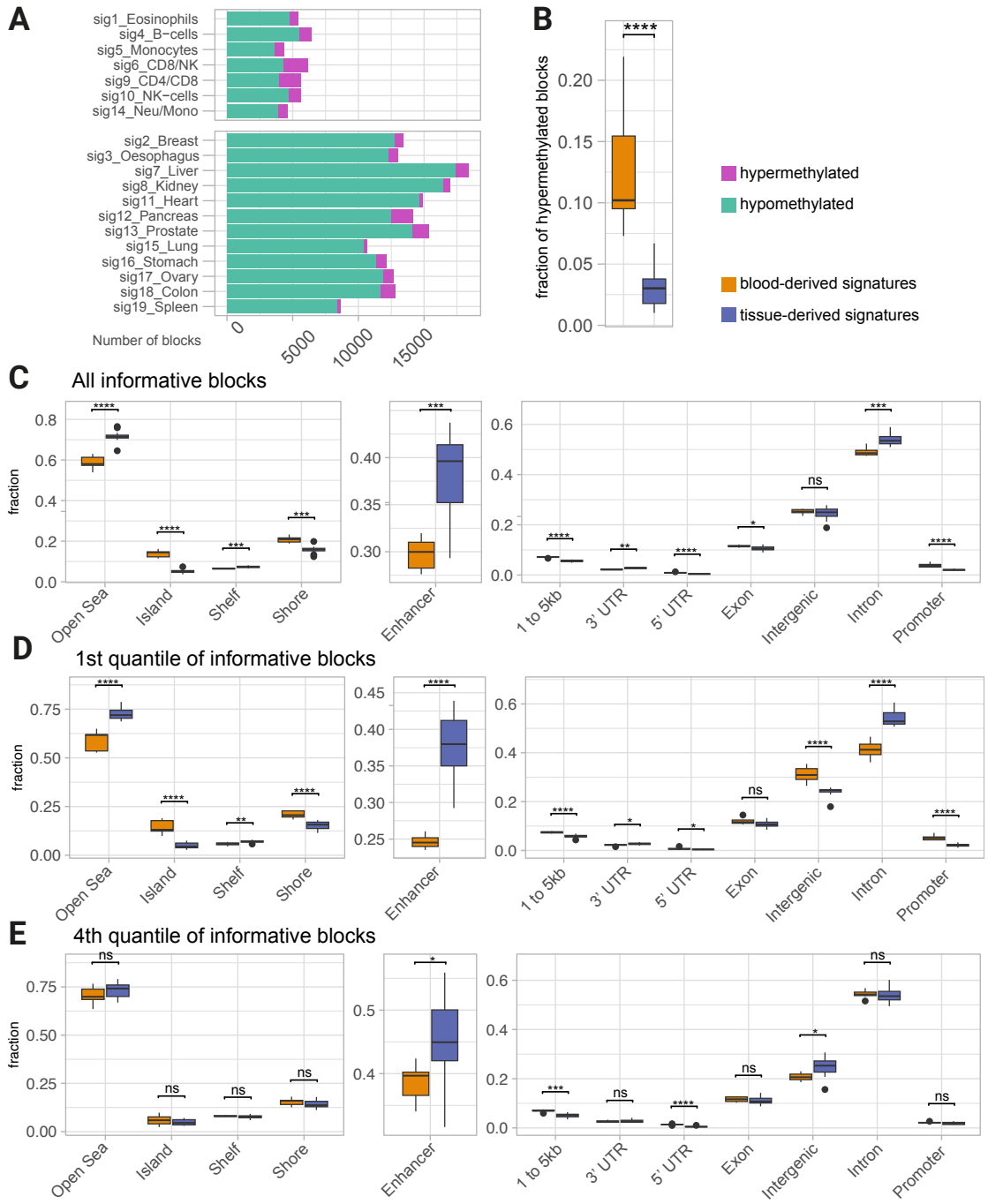


Figure 2.12: Differences in informative blocks between tissue and blood signatures A. Total number of informative blocks per signature, divided into blood-derived signatures (top panel) and tissue-derived signatures (bottom panel). The hypo/hypermethylated block distinction is based on whether the majority of the samples presented a high or low methylation status at the given block. B. Comparison of the fraction of hypermethylated blocks to make the total of the important blocks. Blocks from blood-derived signatures consist of a significantly higher proportion of hypermethylated CpGs, especially signatures 6 and 9. C. Fraction of important blocks that fall within each of the CGI, enhancer and genic elements, separated into blood and tissue-representing blocks. The blocks were further divided into blocks within the 1st quantile of important blocks (based on their coefficient in a given signature) (D) and blocks in the 4th quantile (E). The top-quantile blocks for both blood and tissue occupy genomic regions with a frequency more similar to that of the remaining blocks. The P-values were calculated using the Wilcoxon signed-rank test; * $p<0.05$, ** $p<0.01$, *** $p<0.001$, **** $p<0.0001$.

2.3 Validation

2.3.1 Validation of the selected blocks

Having identified the CpG blocks of interest and their contribution to each tissue-specific signatures, I sought to validate these regions on a new dataset not used in the process development. This check was carried out to confirm that the identified sites are not specific to the particularities of TAPSβ sequencing or the processing of our tissues. In January 2023, a multi-tissue WGBS atlas was published by Loyfer et al. 2023, consisting of 189 samples from 59 tissues. Before bulk sequencing, the samples were FACS sorted to ensure higher homogeneity of the data, which is new, especially for the tissue material.

In order to do this, I downloaded their datasets and processed them in the same way as the TAPS samples, with a few modifications due to the differences in the available data formats, as described in Methods 6.1.2. The full summary of the samples included in the validation is presented in table A.2. I calculated the beta values for all 58,004 blocks and prepared the dataset to fit the modifications applied to the input matrix to the NMF, which included "flipping" the same blocks as in the reference data.

I prepared the new matrix U, consisting of n samples and p CpG blocks, which correspond to the p CpG blocks from the feature matrix W obtained during the NMF on TAPSβ atlas. I used non-negative least squares regression (NNLS) to calculate the contributions of signatures k to samples from matrix U (summarised in Figure 2.13). This approach does not allow for the possibility of any "unknown" tissues being present in the atlas and will always succeed in fitting the new samples to our matrix W, although with a varying value of the fit residuals. For example, if the entry atlas does not contain a liver sample, but the unknown sample to be deconvoluted is derived from liver, NNLS will do its best to assign it to the signatures present in the atlas. It may indicate the closest possible tissue and / or return relatively high residuals.

The results of the NNLS fitting show that most of the samples that had their equivalent in the TAPSβ atlas are correctly matched with the appropriate signatures (Figure 2.14). Due to the lower granularity of our data, samples coming from subsets of cells from a complex tissue are assigned to the signature representing the bigger tissue, for example, all pancreatic subtypes have a high contribution of signature 12. Some signatures, such as signature 3, capture not only oesophageal samples but also closely related tonsil, tongue, and several lung samples,

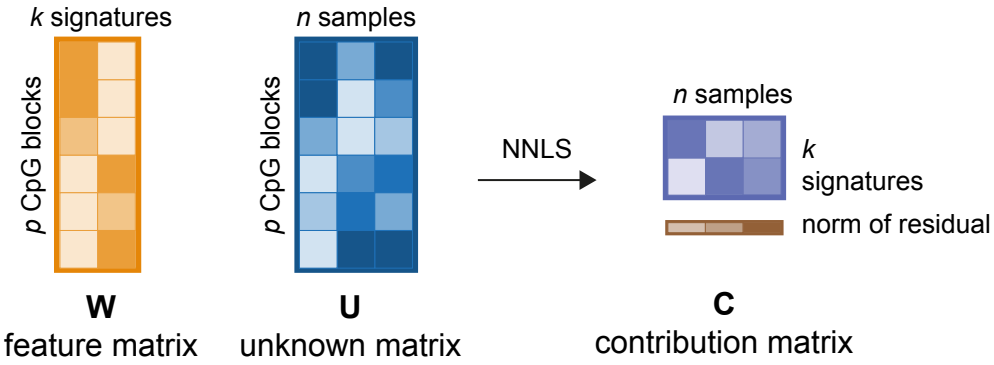


Figure 2.13: Using NMF results to deconvolute unknown samples
Applying NNLS to fit an unknown matrix U to a feature matrix W obtained from *de novo* NMF will result in a new contribution matrix C. This matrix illustrates the contribution of each signature k to each sample n. Additionally, a residual norm is calculated for each sample, illustrating the goodness-of-fit. The intensity of the colour corresponds to the value of the coefficients in W, the state of methylation in the matrix U and the relative contribution of the signature in the matrix C.

potentially encapsulating the general squamous epithelium cells. Samples from peripheral blood nearly perfectly matched their signatures, including cases where the signatures are quite unclear, such as signature 6 and 14, which could not be matched with just one cell type in the *de novo* NMF. Tissue samples which have contributions from more than one signature offer an interesting insight into the nature of the CpG blocks. This is the case when a studied sample does not have the signature equivalent, such as in the case of the thyroid or bone marrow. There is also a visible separation of signals from endothelial and epithelial origins, as illustrated with the kidney sample: The epithelial cells clearly represent the kidney signature, whereas endothelial equivalents are split between two, potentially unrelated signatures. This suggests that despite bulk sequencing of unsorted tissue, there are still some dominant subsets. As noted before, samples without the equivalent signature will still be assigned to other signatures, as there is no option for "unknown" category. The proxy for that is looking at the norm of residuals, providing insight into the goodness-of-fit. It is clearly observable that blood cell types show a much lower sum of residuals, potentially because of the low number of high-weight CpGs blocks contributing to the signature. The tissues matched with tissue signatures show a much higher norm of residuals because of the fitting to many more high-weight CpGs. Expectedly, the cells without the appropriate equivalents in the atlas show the highest norm of residuals, as the model struggled to fit them in the most efficient way.

In conclusion, this result suggests that our method and the identified CpG blocks correctly represent the tissues of origin, and the methylation signatures are due

to the nature of the tissues rather than the type of sequencing or experimental handling. Additionally, the new methylation maps were obtained with WGBS - while small changes may occur, the exclusion of 5hmC in TAPSβ does not greatly affect the tissue-specificity of the produced signatures.

2.3.2 Validation of the genome segmentation and filtering method

The method described in this chapter was developed using the TAPSβ dataset and is summarised in Figure 2.15. To test whether the method can be easily adjusted to new datasets, to incorporate additional samples or test completely new data, I re-examined the process using the dataset published by Loyfer et al. 2023 to test if it is replicable. If it is, I would expect to gain similar results with minimal changes to the method.

The initial data processing was performed as described above and in the Methods section 6.1.2. Due to the unavailability of read-based data, I skipped the SNV removal process. The remaining steps were performed without modifications, and the summary of the samples and tissue grouping used is presented in Supplementary Table A.2. Genome segmentation produces a slightly larger number of blocks than that obtained in the TAPSβ atlas – 998,044. Filtering out low-variance sites resulted in a considerably lower number of blocks, ending at 29,468 for 0.01 cutoff and 50,674 at 0.005 cutoff. Given that 0.01 was the cutoff used in the original method, but the cutoff of 0.005 resulted in a more similar number of blocks, I investigated both scenarios to assess the need for manual cutoff changes in applying the method to new datasets. The methylation landscape of the blocks selected using the 0.01 cutoff is shown in Figure 2.16 (0.005 variance filtering was not shown, because the patterns were nearly identical). There is a striking difference in how cell type-specific the blocks are, specifically due to the lack of multi-tissue regions as seen in Figure 2.4. This is most likely due to the sorting of the cells prior to sequencing, which makes the cells more homogeneous and removes any "contaminating" cell types, such as fibroblasts or residual blood cells from the sequenced samples.

The blocks obtained after filtering overlapped genomic regions in a similar fashion to the blocks in the TAPSβ atlas (Figure 2.17). To test whether relaxing the variance filter enriches for specific genomic regions, I investigated where the blocks unique to the 0.005 cutoff are present. There is a slightly higher preference for intergenic and promoter regions, and a lower fraction of enhancer regions is

Figure 2.14: Deconvolution of new samples with TAPSβ atlas
The normalised contributions of each signature to samples from Loyfer et al. 2023 atlas. See main body for interpretation. Adjacent sample names representing the same cell types were removed for clarity.

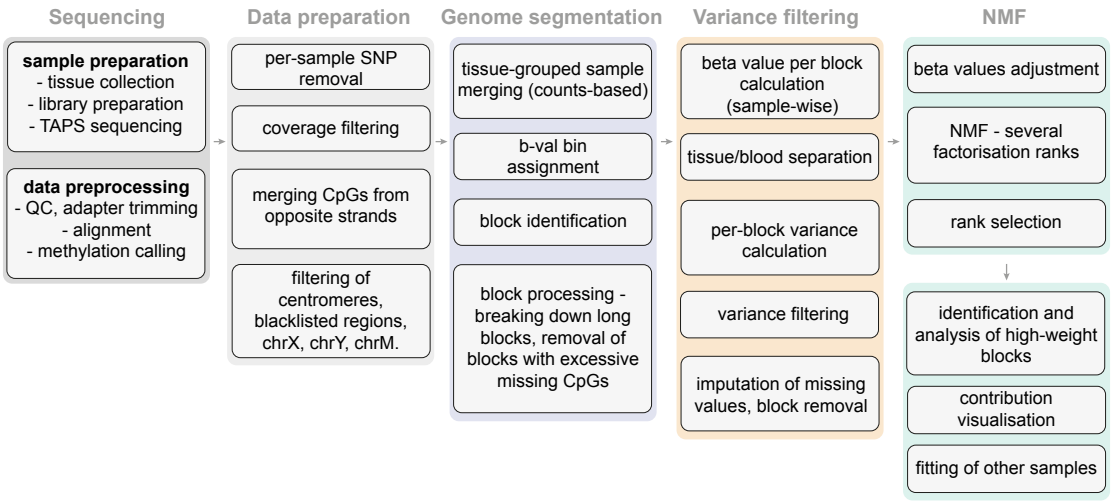


Figure 2.15: Overview of the entire pipeline
Summary of all steps in the process.

selected, although in total, the representation of all regions increases with the larger number of blocks.

I applied the NMF on both sets of blocks, using the same method as described above. Given the larger number of samples, I tested the algorithm on a wider range of possible signatures (k). As illustrated in Figure 2.18, there are several peaks of the cophenetic value that reflect the stability of the signatures obtained. The first peak is at k=22, which is considerably lower than the number of tissue types in the input data, and while it identified several tissue-specific clusters, a large part of the samples were clustered based on the developmental origin rather than tissue (results not shown). The climbing peak towards the right end of the plot represented a k value closer to the number of tissue types; however, it began to split certain tissues into individual samples, as shown previously in Figure 2.10. The middle peak at k = 32 reflected the composition of the atlas in the most balanced way, still failing to assign separate signatures to certain, similar cell types such as different subtypes of gastric epithelium (signature 21), and separating a sample from its main signature in only one case (heart cardiomyocyte, signatures 5 and 27) (Figure 2.19). Applying NMF to the larger set of blocks obtained by filtering on the 0.005 cut-off suggested the same number of signatures and grouped the samples in a nearly-identical fashion. This suggests that the number of blocks is not a factor that determines the output of the pipeline, and the value can be modified depending on the planned downstream analyses.

The coefficients of individual CpG blocks in the identified signatures have a different distribution than observed in the TAPSβ atlas (Figure 2.20). With the

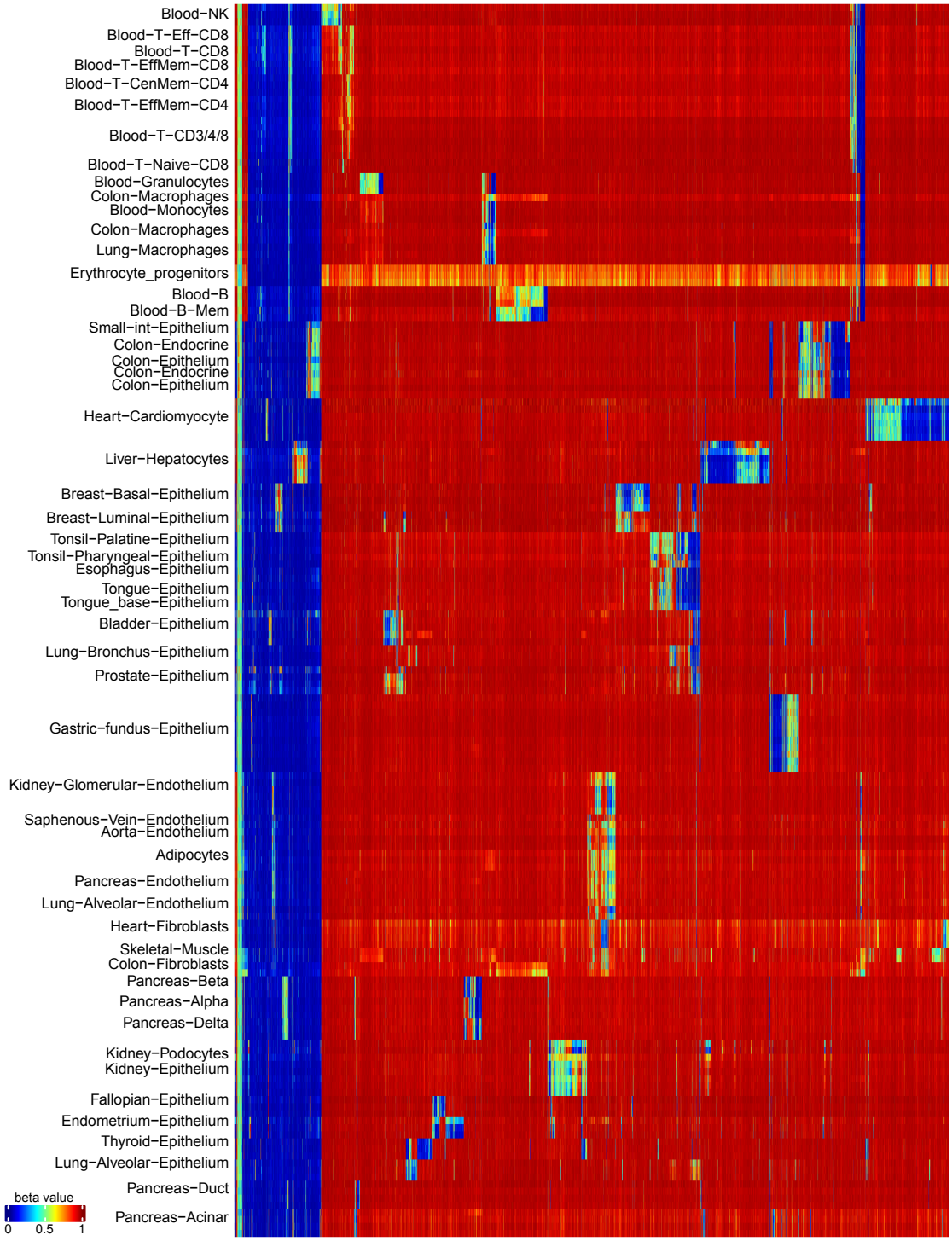


Figure 2.16: Methylation status of highly-variable blocks in the Loyfer atlas
Beta values of blocks with a variance greater than 0.01 per block, computed separately for blood and tissue samples, resulting in a total of 29,468 highly variable blocks. Duplicate names of adjacent samples from the same origin were removed for clarity.

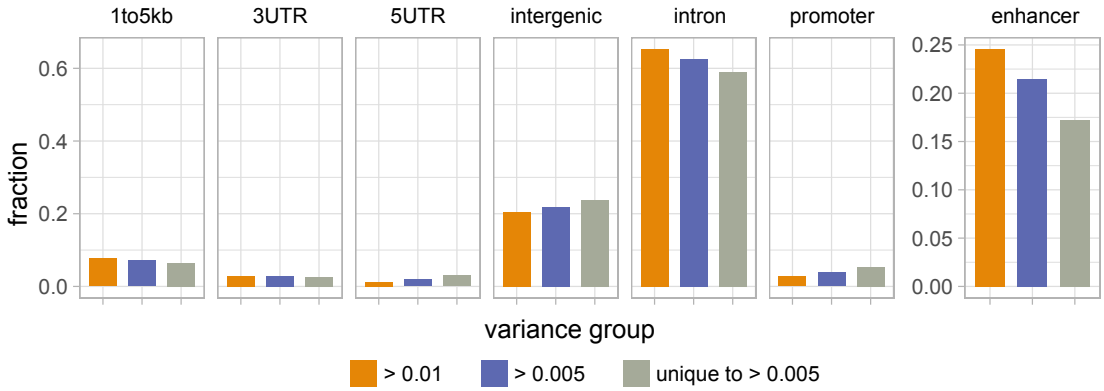


Figure 2.17: Representation of genomic regions
Fraction of blocks that overlap the illustrated genomic regions at two filtering cut-off points. The grey bar illustrates blocks that are unique to the less stringent filtering setting

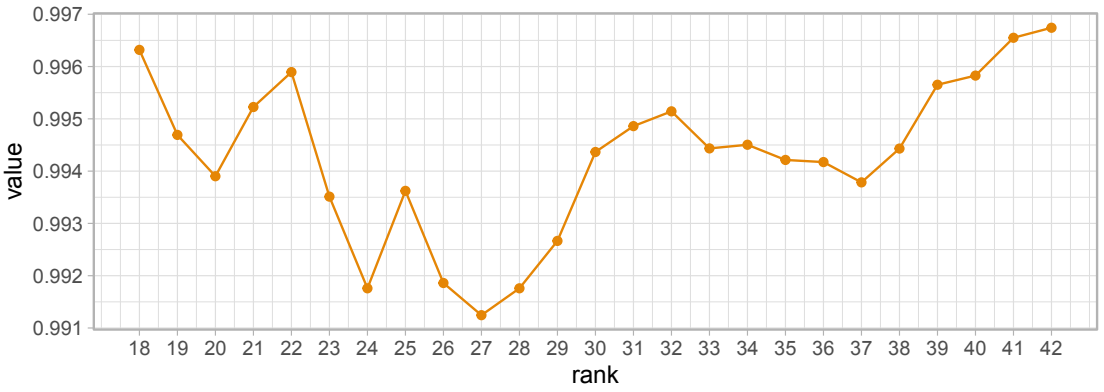


Figure 2.18: Cophenetic values from several NMF runs on the Loyfer atlas

exception of the generally distinct bone marrow-derived erythrocyte progenitors signature, all signatures consist of a major peak and a long tail, similar to the blood-defining signatures in the original atlas. I applied the same method to select the most informative CpG blocks per signature. The resulting numbers were considerably lower than in the TAPSβ atlas, and the differences between the blood and tissue groups were no longer observable (Figure 2.21). Given the similarity of the pattern to the patterns observed in blood samples, it is likely that it has to do with the purity of the included samples, such that there is no mixture of various cell types creating additional, lower-importance blocks per signature.

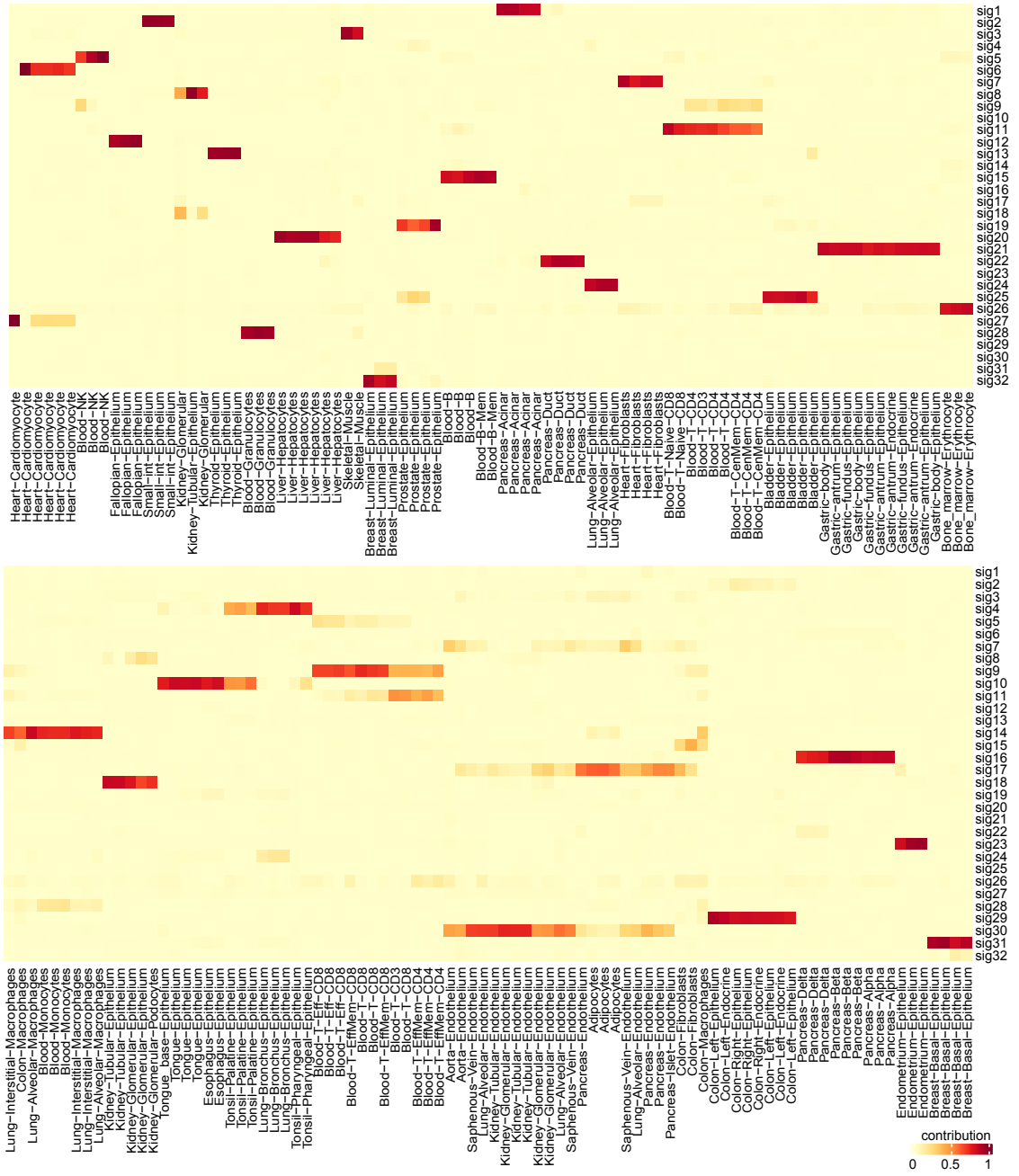


Figure 2.19: De novo signature identification in the Loyfer atlas.
Many unique cell types present in the atlas were assigned a unique signature using a factorisation rank of 32.

Fitting our samples to the Loyfer signatures

As a last piece of validation, I reversed the previous experiment and deconvoluted our TAPSβ atlas samples using the Loyfer-defined CpG blocks and coefficients. I followed the same methods as described in the previous section and used the 32-signature set and the 29,468 high-variance blocks (Figure 2.22). The

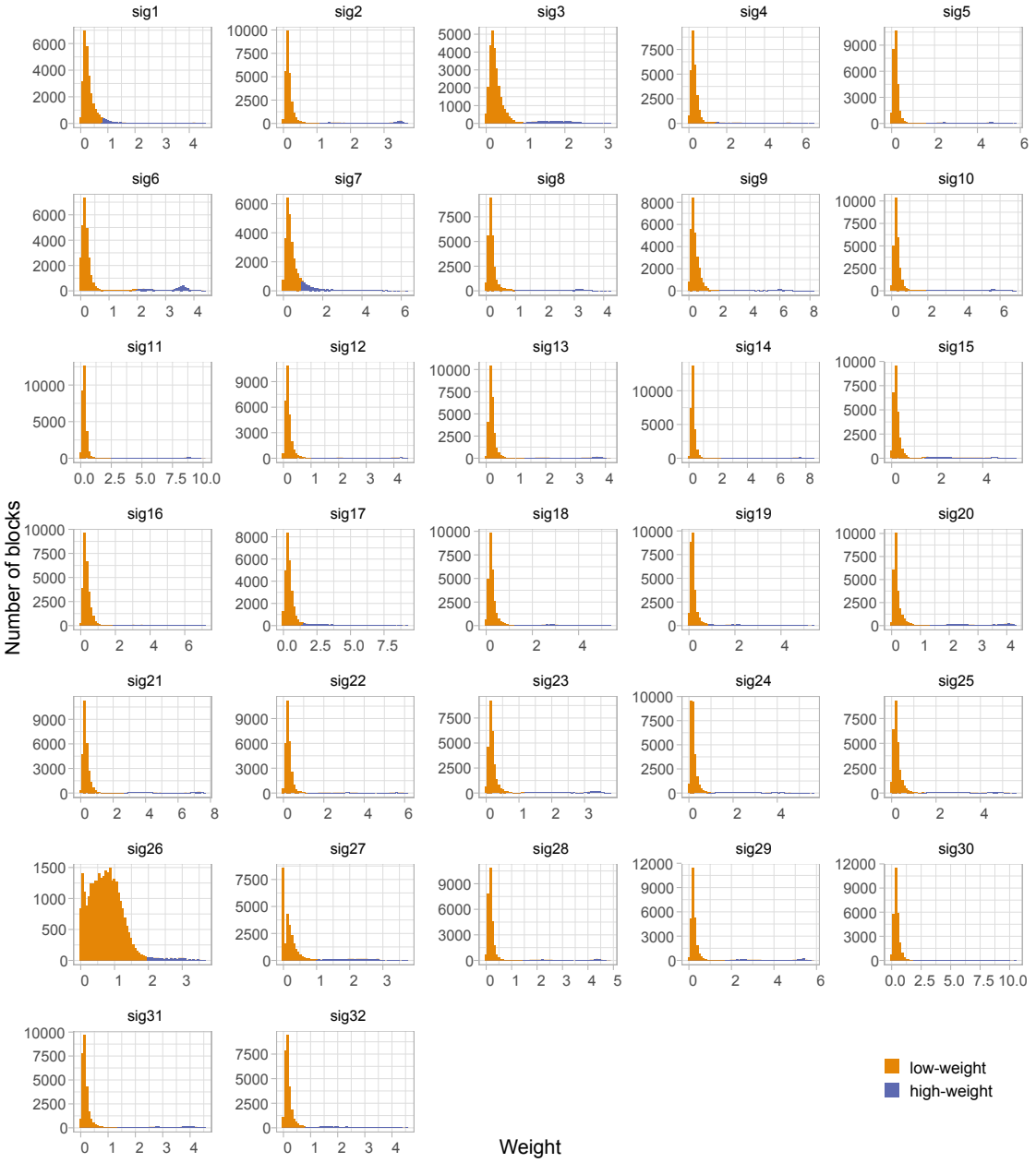


Figure 2.20: Weight distribution varies across Loyfer atlas signatures.
Most of the signatures display weight contributions similar to blood samples from Figure 2.11. The colour of the areas represent the blocks considered informative for each signature.

deconvolution assigned the appropriate signatures in most cases, especially to peripheral blood cells, the fits of which are also accompanied by a low residual value. An exception to this is that of eosinophils, which are not present in the Loyfer dataset, but were correctly assigned to other cell types of a similar developmental origin. The tissue samples, while with high contributions of

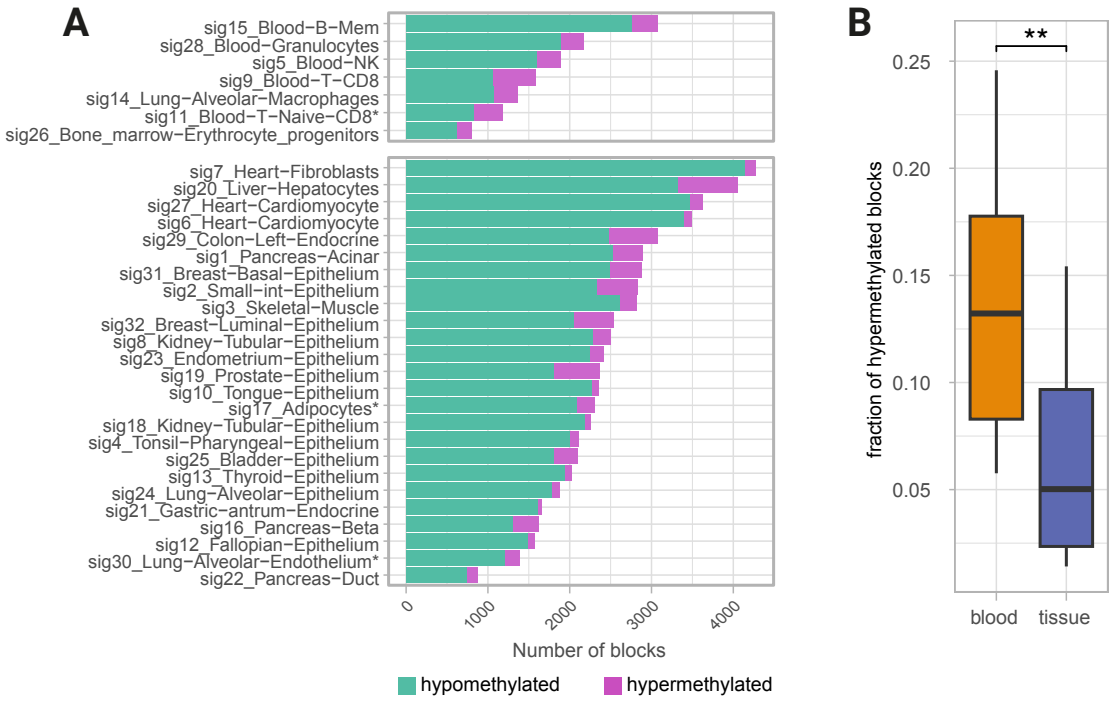


Figure 2.21: The differences in numbers of selected blocks from tissue- and blood-derived samples
A. The total number of informative CpG blocks, divided by blood and tissue samples, with highlighted hypo and hyper-methylated subgroups. B. Combined difference in block fraction The p-value was calculated using the Wilcoxon signed-rank test; * $p<0.05$, ** $p<0.01$, *** $p<0.001$, **** $p<0.0001$.

appropriate signatures, have a noticeably lower contribution of those signatures on a whole, diluted by signals from other signatures, such as signature 17 (adipocytes/epithelium) and signature 26 (bone marrow), which capture multiple cell types or different methylation landscapes (bone marrow).

Overlap of the identified blocks

Loyfer et al. 2023 has also identified several tissue-unique CpG blocks, and I compared them to the sets found using my method. Although I used their most generous list of top 1000 blocks per tissue type that covers 197,901 CpGs in total, the CpGs obtained using our method had a very small degree of overlap of 9.7%. This is somewhat unexpected given the same input data used for the block identification. Loyfer cell types with the highest overlap with our method included samples such as heart and thyroid (20-25% of Loyfer sites present atlas curated with our method), and the smallest overlap was for samples that did not separate into their own signature, such as colon fibroblasts (0.03% shared CpGs). This is most likely due to less supervised nature of our method, as I do not "force"

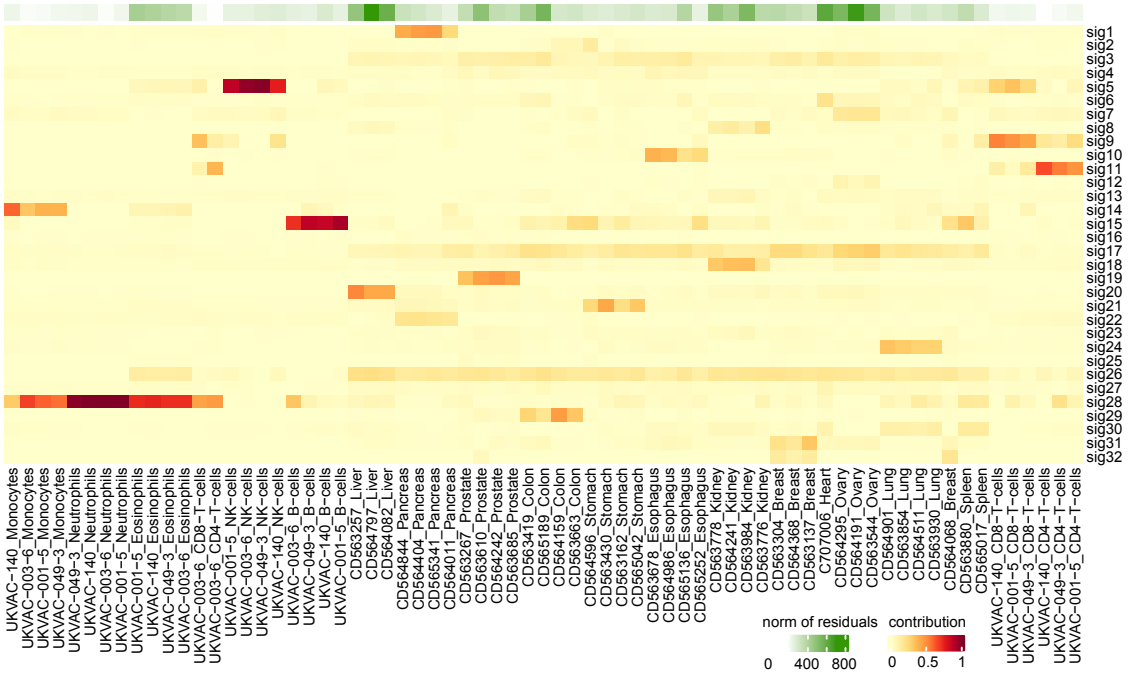


Figure 2.22: NNLS fit of Loyfer atlas signatures to the TAPSβ samples

a certain number of tissue-specific blocks. Additionally, there was also a relatively small overlap with the CpGs identified in the TAPS atlas, which is probably caused by the presence of different tissues in the dataset, as well as using the different method (Figure 2.23).

2.4 Discussion

In summary, I have developed a method allowing the identification of highly variable, tissue-specific CpG blocks across the genome. The blocks span both the genic, intergenic, and regulatory regions of the genome, and allow for the identification of sites most specific for a given tissue, while allowing for certain lineage-specific similarities. I applied the method to heterogeneous data, sorted cells, and to a mixture of both, and I found it to require minimal optimisation to be applicable to new datasets.

The method has been developed on TAPSβ methylation calls, which indicates the levels of 5mC, but not 5hmC which is included in sequencing using other technologies, such as WGBS. I did not notice any substantial issues with the application of the method on bisulfite and bisulfite-free sequencing, and I hypothesise that due to the low content of 5hmC in the analysed cells, the incorporation of CAPS data would not affect the method development process. Caution should be

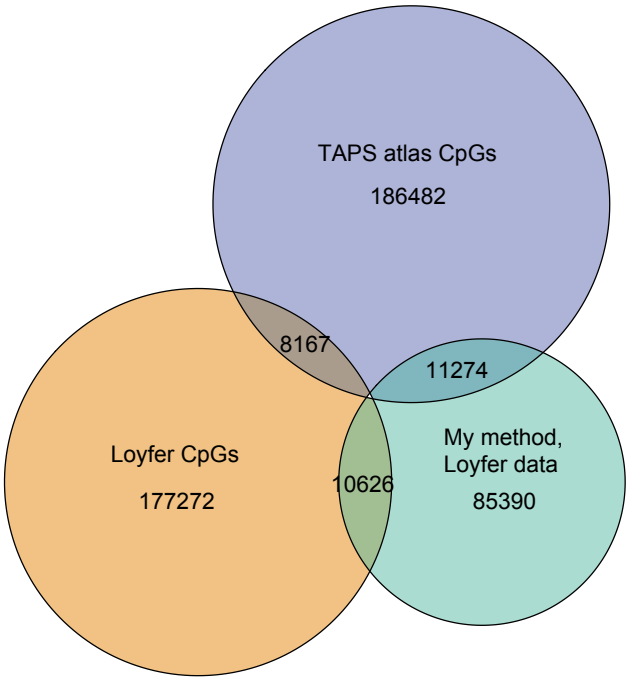


Figure 2.23: Overlap of CpG blocks in Loyfer data and TAPSβ samples

applied when analysing brain tissues, where 5hmC was shown to have different properties and distribution than in the rest of the genome (further discussed in main discussion) (Kriaucionis and Heintz 2009; Globisch et al. 2010; Guo et al. 2011; Szulwach et al. 2011; He et al. 2021).

The segmentation of genome prior to the site selection and deconvolution greatly simplified the subsequent analysis. Similar approach was previously used by Guo et al. 2017, who found the regional methylation patterns can be represented as highly-coordinated blocks. More recently, the deconvolution study by Loyfer et al. 2023 developed a similar method, identifying change-points in correlated methylation values and slicing the genome accordingly. Both of these methods were read-based, allowing for fine-tuning of the identification constraints. The initial part of the development of my approach was limited to the available public datasets from ENCODE and Roadmap consortia and the available .bed files. Despite the limitations, I obtained a large number of blocks covering $\tilde{2}$0% of all CpGs, and selected highly-variable blocks using a non-supervised approach. Apart from the variance cut-off points, I set no restrictions to the number of blocks "representing" each tissue. This is in contrast to the majority of reference-based deconvolution studies published to date, which tend to choose a small number of most unique blocks or single CpGs, imposing an equal number of sites representing each tissue (Loyfer et al. 2023; Moss et al. 2018; Guo et al. 2011;

Chakravarthy et al. 2018). While the strict selection of top blocks is necessary to achieve the highest-possible specificity in downstream applications such as cfDNA signal detection, it removes large quantities of potentially interesting biological information. The details of the segmentation and filtering can be adjusted, as demonstrated with the separate filtering of blood and tissue samples to retain higher numbers of blocks. The genomic distribution of high variance blocks reflects the expected locations of tissue-specific CpGs, such as gene bodies, distal regulatory regions and locations adjacent to CGIs. Both CGIs and promoters were strongly selected against, despite the high over-representation of these regions after genome segmentation, caused by their uniform demethylated state, often included in the blocks.

NMF in the context of methylation has been used in the reference-free deconvolution of immune cells (Houseman et al. 2016), but not to construct the reference atlas to be used in the deconvolution of further samples. I demonstrated the applicability of NMF in this context, assigning appropriate weights to CpG blocks representing each group of tissues included in the atlas. A more in-depth analysis of the assigned coefficients is described in the next chapter.

I applied this framework to both heterogeneous, bulk tissue-derived methylation maps, FACS-sorted individual cell types, and a mixture of both, with satisfying results obtained in cross-validation. Despite potentially containing a range of cell types, the bulk tissue heterogeneity may not be a "confounder"; in fact, it appreciates how complex these tissues are. By definition, sorting and filtering cells will remove certain cells from the tissue. As much as we know about the samples, we are currently unable to isolate and sequence every cell type present within them. This results in the removal of some cells, and the methylation patterns obtained would represent an incomplete landscape of the sample. Bulk sequencing, in contrast, will retain all cells from the sample, and the total methylation landscape reflects the entirety of the mixture. Of course, this may lead to noise and multiple tissues sharing the same blocks, but the noise can be removed or accounted for in further analyses. Depending on the research questions, both approaches have reasonable use cases. The surprisingly small degree of overlap between the CpGs included in methylation arrays and in this atlas highlights the benefits of developing whole-genome based approaches in increasing the numbers of identified biomarkers. Similar observations have also been noted by Loyfer et al. 2023.

The method offers a simple way to segment the genome and identify the most variable sites, corresponding to tissue-specific sites. In the next chapter, I will further explore the topography of these sites and the characteristics of which particular aspects of the placement play a role in tissue specificity.

3

Functional analysis of signature-specific blocks

Contents

In the previous chapter, I identified nearly 60,000 CpG blocks with tissue-specific methylation patterns. In this section, I investigate why these specific blocks are likely to have these properties, by analysing the methylation and hydroxymethylation patterns of the selected blocks together with over-representation analyses and introducing higher-order epigenetic modification data.

3.1 Patterns of coefficient weights across selected blocks

Before attempting to perform functional analyses on the blocks identified in previous chapter, I investigated the patterns of coefficient weights assigned to each block in every signature. As hinted in the previous chapter, the distribution of weights varies between the signatures. To gain a better understanding of these

distributions and investigate how NMF handled the selection of blocks defining each signature, I created a heatmap of the NMF-assigned weights (Figure 3.1 A). For clarity, all weights were scaled to 1, and blocks which did not satisfy the signature-wise cutoff were replaced with 0s. I identified three major clusters using k-means clustering. The majority of the selected blocks belong to the second cluster, representing blocks mostly unique to each signature. The second largest cluster contains blocks that present high weights in only signatures representing solid tissues. Lastly, the third group consists of CpG blocks with the strongest signal coming from blood-derived samples. However, despite the clear signal, the remaining signatures also contain many blocks above the cutoff point in this cluster. While clearly distinguishable when looking at relative weights, the distinction is lost when we are only looking at the "binary" outcome of whether or not the particular block is included (Figure 3.1 B). Additionally, I plotted the corresponding average 5mC levels at each block from tissues with highest contributions of each signature (Figure 3.1 C). It is clearly visible that the patterns in the 5mC map correspond nearly perfectly to the patterns of high-weight blocks in the previous panels.

Although these distinct distributions of weights should not have an impact on the NNLS fitting of other datasets, the peculiar patterns in the first and last clusters should be taken into account in the functional analyses, especially when using the binary above-cutoff classification rather than direct weights.

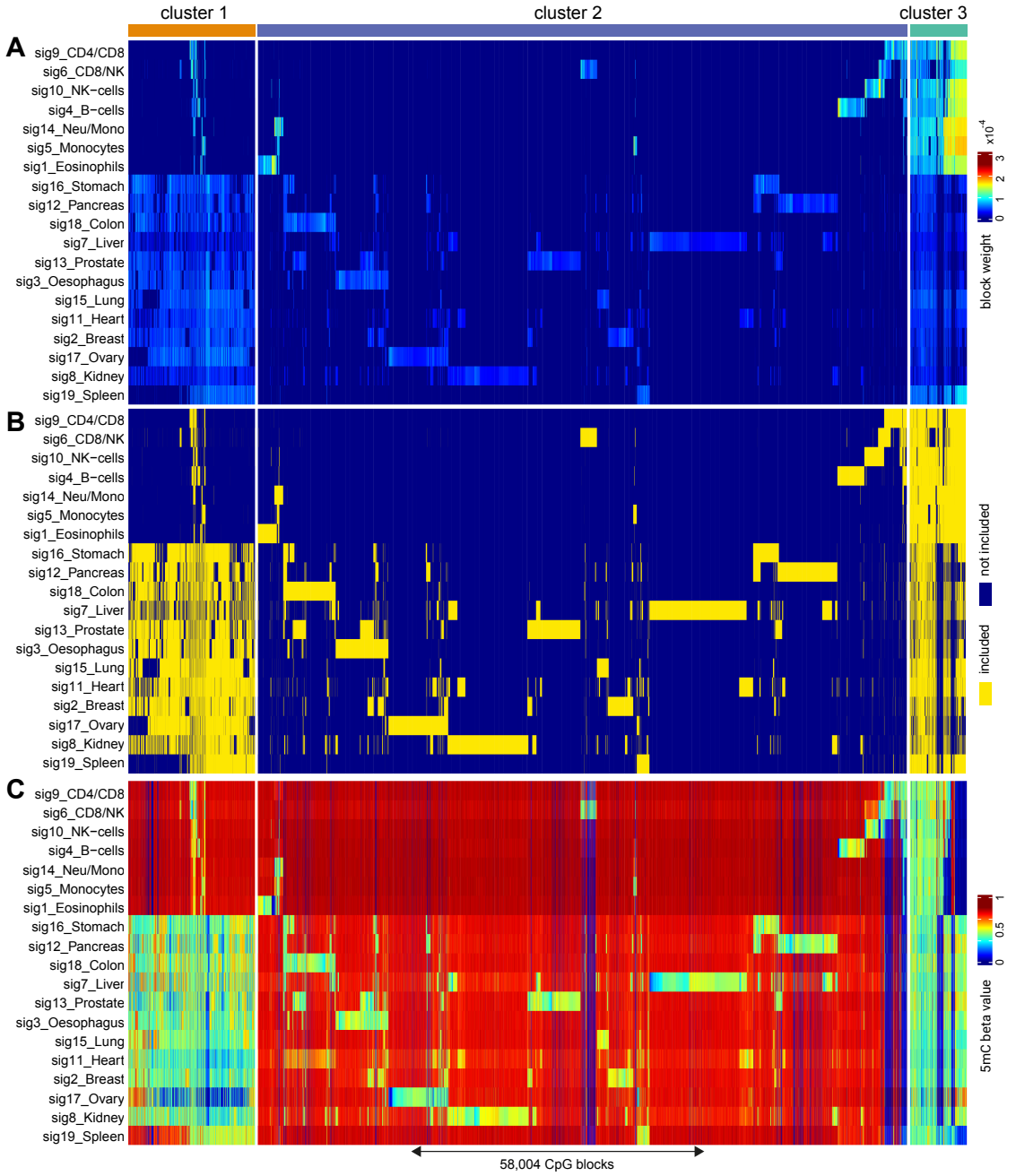


Figure 3.1: Weights of signature-defining blocks
The weight of the blocks normalised to 1 with any non-important blocks replaced with 0s. A. Relative weights of blocks, B. Binary values based on being classified as included (above cutoff point) or not. C. Average methylation value of the blocks calculated using samples with the highest contribution of the corresponding signatures. The rows and columns are ordered according to the k-means clustering of the matrix in panel A.

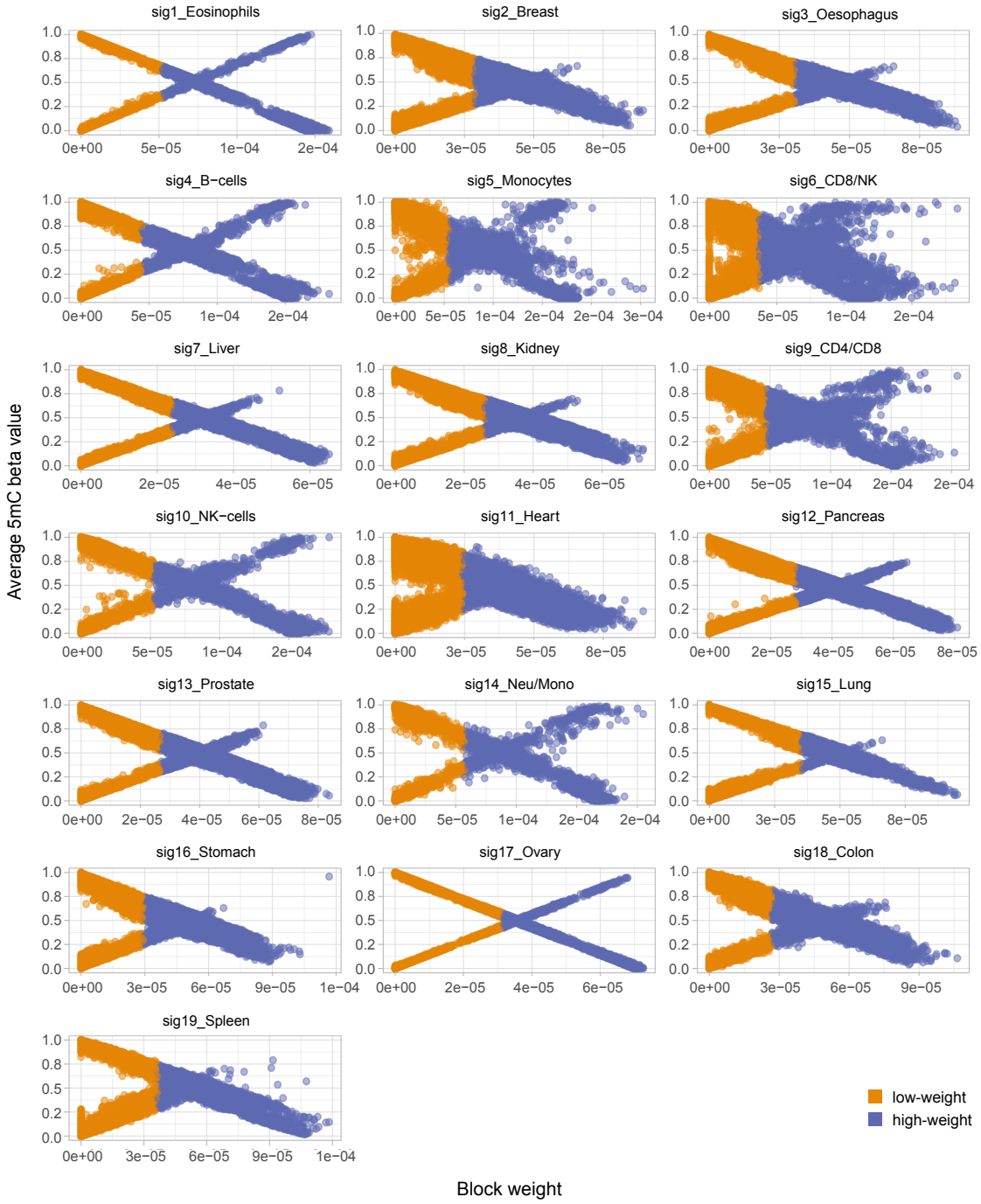


Figure 3.2: Correlation between beta values and block weights
The average beta values per block were calculated from samples with the highest contribution of the relevant signatures and plotted against their coefficient weight assigned by NMF.

3.1.1 Correlation of 5mC levels with block weights

To better understand the relationship between the NMF-assigned weights and the observed 5mC levels, I plotted the average 5mC beta values against the weights

in each of the signatures (Figure 3.2). Almost every facet contains an X-shaped formation of data points. The purple-coloured points mark the blocks which were selected as important in the previous cutoff selection. The clarity of the X-shape comes from the algorithm behind NMF, where the highest weight is assigned to the most distinct sites, which in our case are the most hyper- or hypomethylated blocks relative to the beta values in the remaining samples. The lack of the positively-correlated diagonal line in a couple of facets is associated with signatures with very low numbers of hypermethylated blocks, as seen in Figure 2.12 A. The distinct shape of the points has to be taken into consideration when performing the functional analyses, as the weights do not necessarily represent the strongest biological impact, but rather the most distinct values. For this reason, in the following analyses, I consider the blocks in a binary fashion, where each point either is, or is not, important for each signature. This ensures that the analyses are not biased towards the more extreme methylation values.

3.1.2 Functional analysis of the high-weight blocks

Genomic representation of the three clusters

To begin the functional analysis of the groups of high-weight blocks, I investigated whether there are differences in the general genomic representation of the blocks belonging to the three identified clusters (Figure 3.3). As indicated above, the first cluster is generally multi-tissue specific, the second cluster represents the signature-unique blocks, and the third cluster represents blocks of highest weight in blood cell types (i.e., highly hypomethylated in blood cell types) mixed with blocks of high weight from the remaining signatures. I overlapped the locations of blocks in each cluster with various genomic locations similarly to the previous analyses and found several differences between the clusters. The first cluster is remarkably rich in distal enhancers, at a rate of 20% more than all high-variance blocks on average. The second cluster represents values most similar to the average high-variance blocks as presented before in Figure 2.6. Cluster 3 is relatively low in introns and open-sea regions, higher in promoters, and slightly elevated in enhancers.

Over-representation analyses

There are several approaches to functional analysis of genomic regions. All rely on the identification of genes overlapped by our regions of interest, which are then tested against predefined sets of genes representing biological processes,

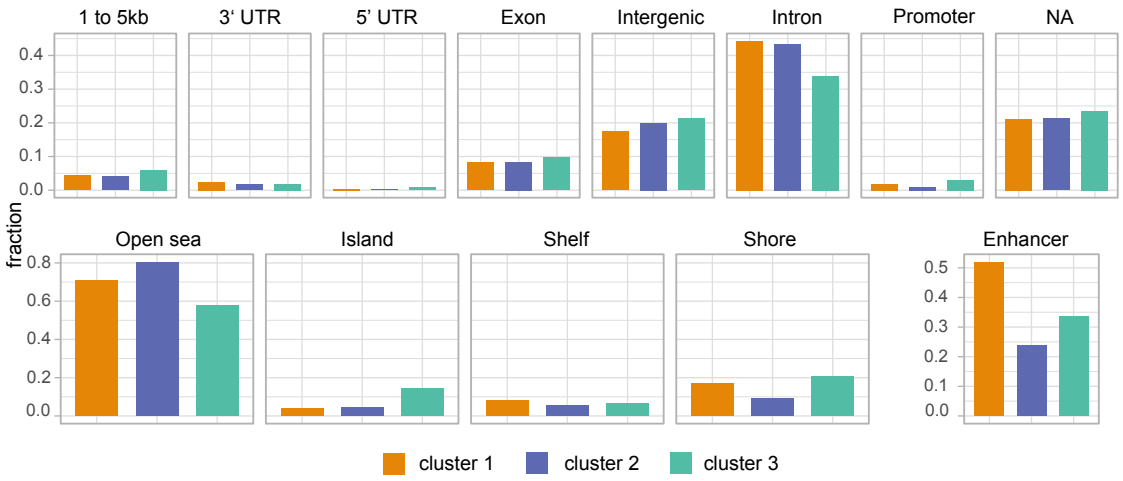


Figure 3.3: Comparison of the distribution of the clusters across the genome
The clusters correspond to 1) multi-tissue methylation signal, 2) tissue and cell-type specific blocks, and 3) predominantly immune cell-specific blocks.

pathways, or various other annotations. Based on the characteristics of the input data, there are different statistical methods to perform the analyses to determine the enriched gene sets. In the case of my data, all analyses presented below are conducted using over-representation analysis (ORA), which relies on performing multiple hypergeometric tests between the input genes and reference gene sets (Khatri et al. 2012). ORA does not rely on weights or order of genes, just on their presence or absence in the test sets. In order to find genes associated with CpG blocks, I overlapped the genomic locations of blocks with the locations of genes in the The Matched Annotation from the NCBI and EMBL-EBI (MANE) transcript databases (Morales et al. 2022). Apart from gene bodies, I also included the proximal regulatory elements, such as promoters and the regions upstream of promoters, widening the possibility of being included in the gene set.

Firstly, I hypothesised that the genes overlapped in cluster 1 may be representing some form of solid tissue-specific functions, while cluster 3 would be more blood-specific. To test this, I performed ORA on the Gene Ontology (GO) Biological Processes (BP) gene sets, separately for each of the clusters (Figure 3.4). Cluster 2 returned no significant hits, and I removed any hits that were shared by both cluster 1 and cluster 3 for clearer interpretation. There are no obvious differences between the clusters suggesting the involvement of tissue or blood-specific processes, although several high-scoring hits are of interest. For example, cluster 1's most significant sets is actin cytoskeleton/filament organisation, regulation of cell morphogenesis, and several hits relating to blood vessel formation and

angiogenesis. It is challenging to interpret these results, but they may suggest the involvement of these blocks in general development and cell growth. On the other hand, cluster 3's highest-scoring sets relate to cell migration, and there is a whole group in neuronal sets such as synaptic signalling and neuron-fate commitment. These results are even more perplexing to interpret, but suggest taking the results from these two clusters with caution. Testing other GO gene sets and KEGG pathways resulted in even more inconclusive results or very few hits.

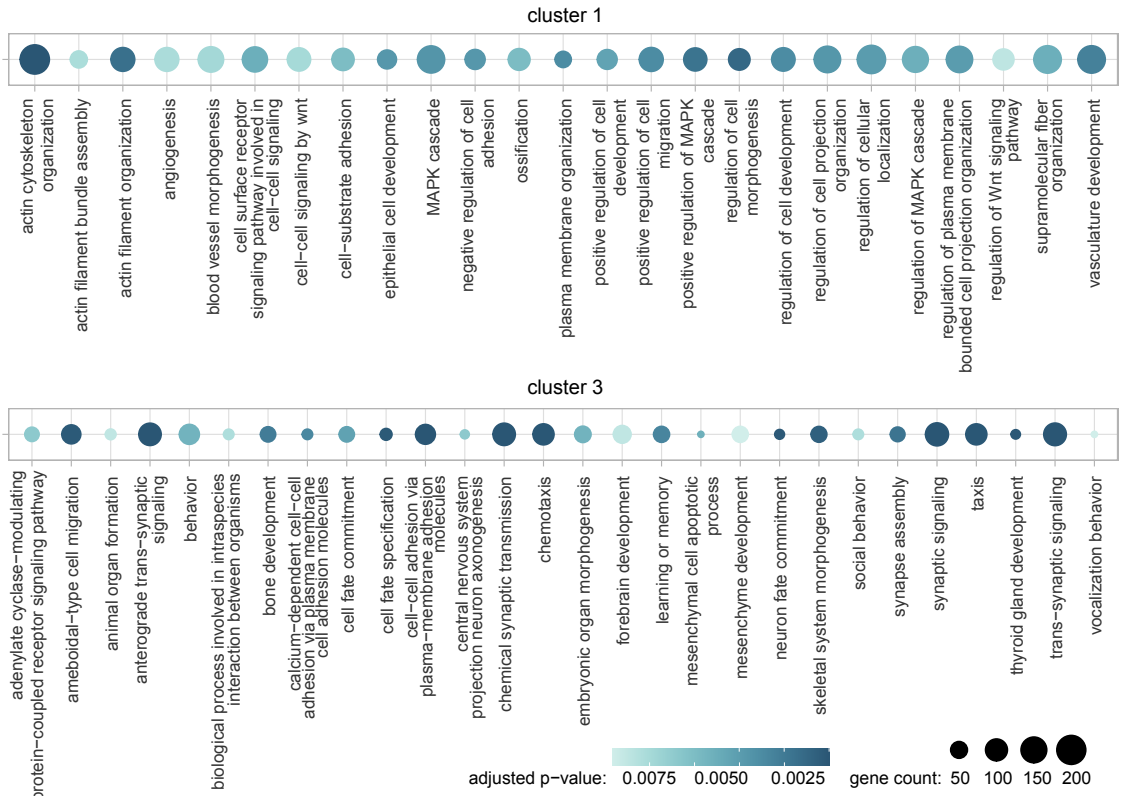


Figure 3.4: ORA on genes covered by clusters 1 and 3
The dot plot illustrates the over-representation of GO BP gene sets in both clusters after the exclusion of terms overlapped by both. Benjamini-Hochberg correction was applied to adjust the p-values.

Gene-based ORA

Having investigated the biological processes associated with the blocks in each of these clusters, I focused on the gene-related tissue specificity of the high-weight blocks. To do this, I performed signature-wise ORAs in several configurations (Figure 3.5). Firstly, for each signature, I overlapped each high weight-block with the genes as described above. Then, I tested each resulting gene list against

several gene sets obtained from The Human Protein Atlas (THPA), which contains genes specific to multiple human tissues and blood cell types (Karlsson et al. 2021; *The Human Protein Atlas* 2023). I performed ORA four times for each signature - firstly for blocks from the entire matrix (upper panel) and then separately for each cluster (remaining panels).

The results of the tissue-specific ORA are somewhat complex to decipher. Firstly, we see that when taking into consideration the high-weight CpG blocks from all clusters, we get a relatively good match between signatures and their expected matching tissues (note that not every tissue present in my atlas is present in the THPA gene sets, such as eosinophils). There is a considerable amount of "noise", illustrated by signals such as monocytes, pituitary gland or bone marrow, enriched in nearly all blood-derived signatures. When looking at the cluster 1 alone, very few signatures get any match, apart from, surprisingly, B-cells. Cluster 2, as expected, gave us the clearest representation of tissue-specific matches. Most signatures are correctly assigned to the corresponding tissue, although usually with a few extra significant matches, which sometimes can be explained by a shared developmental origin or function (for example, endometrium and ovary both enriched in signature 17), but is more challenging in other cases - such as the range of matches to signature 2. One can select different p-value cutoffs to "calibrate" the results, but in some cases this may result in the exclusion of the correct matches.

The results of the analysis of cluster 3 are perplexing, but give interesting insight into the previous findings. Firstly, nearly all signatures have strong enrichment of bone marrow signal. It is worth remembering here that almost entire cluster 3 was "high weight" for almost all signatures, but in terms of relative weights it was clearly more blood signature specific. This can explain the high bone marrow signal, which gets "diluted" with other important blocks in matrix-wide analysis in the case of tissues, but remains a strong component for blood types. A similar signal is present for the choroid plexus and pituitary glands. While the choroid plexus is a place where brain-associated immune cells reside, the enrichment of the pituitary gland is harder to explain. It is interesting whether the association of these two tissues with the brain is perhaps responsible for the number of neuron-related hits for cluster 3 in the previous figure.

Additionally, because of the high contribution of distal enhancers in the whole block population, I wondered whether the addition of enhancer information may influence the results. In this analysis, I selected enhancers with high-confidence

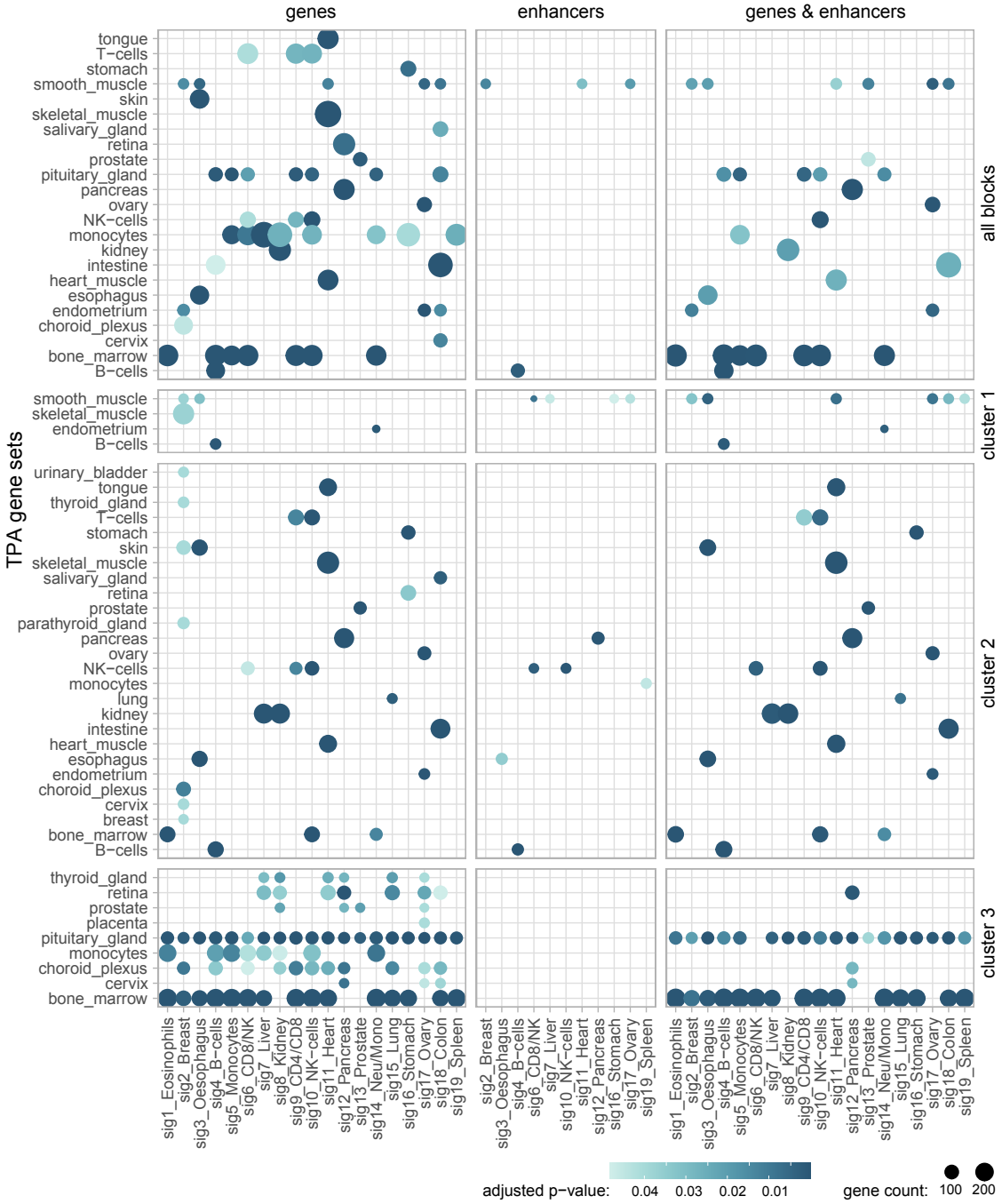


Figure 3.5: Signature-wise ORA on tissue-specific gene sets
Genes overlapping blocks from the entire matrix (top) and each of the three clusters (bottom) in each signature were used as input to ORA using tissue-specific gene sets obtained from THPA. In the first column, only genes directly overlapped by blocks were included. Second column represents genes regulated by enhancers overlapped by CpG blocks, and third column combined both analyses. Benjamini-Hochberg correction was applied to adjust the p-values.

association with genes determined in the GeneHancer database and overlapped the CpG blocks with enhancer locations. Then, for each overlapped enhancer, I created a list of associated genes and used that as input to ORA as in the direct gene analysis above. I tested the enhancer-based ORA both completely separately from the genes (second column) and as a potential to add more information to the genes (third column).

As seen in the second column, the analysis of enhancers using the gene association proxy did not produce meaningful results for most tissues. However, the matches obtained in the analysis of cluster 2 were the most specific of all the analysis, producing very few but well-matched results. The addition of enhancer data to the gene-based analysis had a generally good impact on the clarity of the resulting matches. This may be due to added relevant information in certain tissues, but in the remaining tissues, it is quite possible that the increased test gene sets lead to higher p-values, pushing the less significant hits out of the scale, without necessarily improving the actual signal.

Enhancer-based ORA

The gene-wise analysis provided very good insight into the tissue specificity of the blocks, especially of the second cluster. However, the indirect enhancer analysis did not add a lot of extra information despite the distal enhancers being so enriched in the high-variance blocks in general. This was probably due to the use of an enhancer-gene proxy, which is based on confidence in the enhancers locations and the specific genes they regulate. This can become complicated in cases where enhancers have multiple gene targets, potentially making the results redundant. An alternative approach is to eliminate the need to for the gene association and use databases which directly link enhancers with tissues they are active in. There are several sets included in the GeneHancer database, and for this analysis, I focused on the ENCODE-curated and dbSUPER-curated datasets. The first one deals with normal enhancers, whereas the second focuses on groups of superenhancers, and has a slightly different set of reference tissues. I performed the enrichment analysis in the same fashion as above, but this time, I removed the matrix-wide results due to the noise making the interpretation impossible.

In Figure 3.6, we can see that the first cluster received multiple hits in all tissue signatures, matching a variety of reference enhancer sets. The presence of these hits explains the high representation of enhancers in this cluster. It is quite challenging to interpret this landscape, but it may perhaps be targeting some

Figure 3.6: Signature-wise ORA on ENCODE enhancer-tissue sets
CpG blocks overlapping enhancers included in the ENCODE reference of tissue-specific enhancers were subjected to ORA as described previously. Benjamini-Hochberg correction was applied to adjust the p-values.

housekeeping enhancer network; it is worth noting that despite lower gene counts, multiple hits are significant for also blood signatures. Similarly to gene-based analysis, cluster 2 had the most signature-specific landscapes and correctly matched multiple signatures to cells of origin or similar tissues. There is a considerably higher amount of noise and mismatched samples than in gene-wise analysis. Cluster 3 very clearly shows blood-specific matches, with the addition of muscle and skin signals for the blood signatures. Overall, the enhancer-based analyses provided some insights into the nature of the selected blocks. Neither gene- or enhancer-based analysis yields perfect matches, and this was expected - it is very likely that these signals are complimentary to each other and to other modifications (described below), and together form a truly tissue-specific landscape.

ChIP/DNAse-seq based analysis

Tissue-specific regions are likely to be present in regions of DNA that are accessible to transcriptional machinery. To confirm this, I tested whether high-weight blocks tend to co-occur with DNAse-accessible regions from publicly available DNAse-seq peaks data from the Roadmap Epigenomics Consortium. I overlapped the regions of the peaks with the CpG blocks, and performed Fisher's exact test on the counts of overlaps of high-weight blocks and DNAse peaks from samples with matching Roadmap tissues. In each case, the peaks were significantly over-represented in high-weight blocks, confirming that tissue-specific methylation marks are present mainly in accessible regions (Figure 3.7).

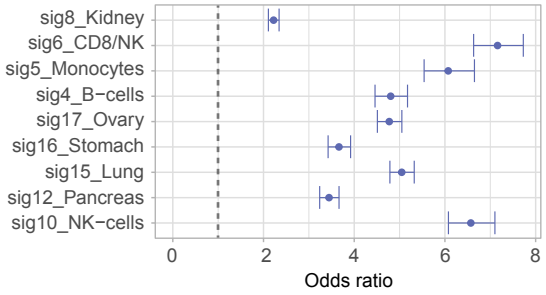


Figure 3.7: DNAse peaks enrichment in high-weight blocks
Blocks from signatures with available Roadmap Epigenomics data were overlapped with DNAse-seq peaks, and their enrichment was calculated using Fisher's exact test. All p values <0.001.

To further assess the tissue specificity of selected blocks, I analysed how well they correspond to ChIP-seq peaks from the same tissues. Tissue specificity of selected chromatin marks, especially H3K4me1, has been described before (Kundaje et al. 2015), so I expected that the high-weight blocks for each tissue overlap with the chromatin peaks of the matched sample to a greater extent than in other samples. I used publicly available data from the Roadmap Epigenomics project and calculated the number of high-weight blocks in each signature that overlap the chromatin mark peaks of each downloaded tissue (Methods 6.5). I performed the chi-square test on each contingency table and plotted the results as a heatmap (Figure 3.8).

As shown in the figure, the blocks that carry high coefficient weights in each signature tend to be over-represented in H3K4me1 peaks coming from the same tissue. Most signatures have a clear match, with the exception of the prostate and kidney signatures, due to the lack of matching samples in the available ChIP-seq datasets. Some tissues such as the lung show higher signals in the foetal tissue,

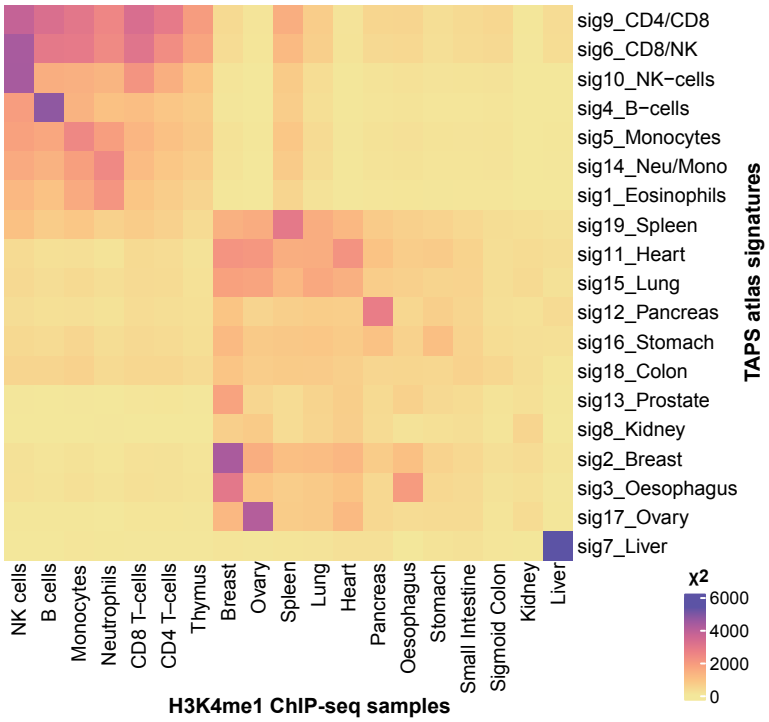


Figure 3.8: H3K4me1 peak enrichment aross signatures
In most cases, important blocks in each signature corresponds to a higher extent to H3K4me1 peaks from the same tissue. All chi-squared values above 0 have the p value of <0.001.

rather than the adult tissue equivalent. The results are distinctly clustered into two groups representing blood and solid tissues. CD4+ and CD8+ T-cells did not represent a clear over-representation of the peaks of any particular blood cell, seemingly representing the blood lineages as a whole more than individual cells. The remaining chromatin marks had minimal tissue specificity.

Given the fact that a large proportion of the selected blocks are not tissue-unique but rather shared across several signatures, I tested whether any of the chromatin marks are particularly enriched in unique or shared blocks. First, I paired each signature with the corresponding tissue from the Roadmap atlas, guided by the origin of the tissue and the representation of H3K4me1 in Figure 3.8, and marked the overlaps with the chromatin mark peaks.

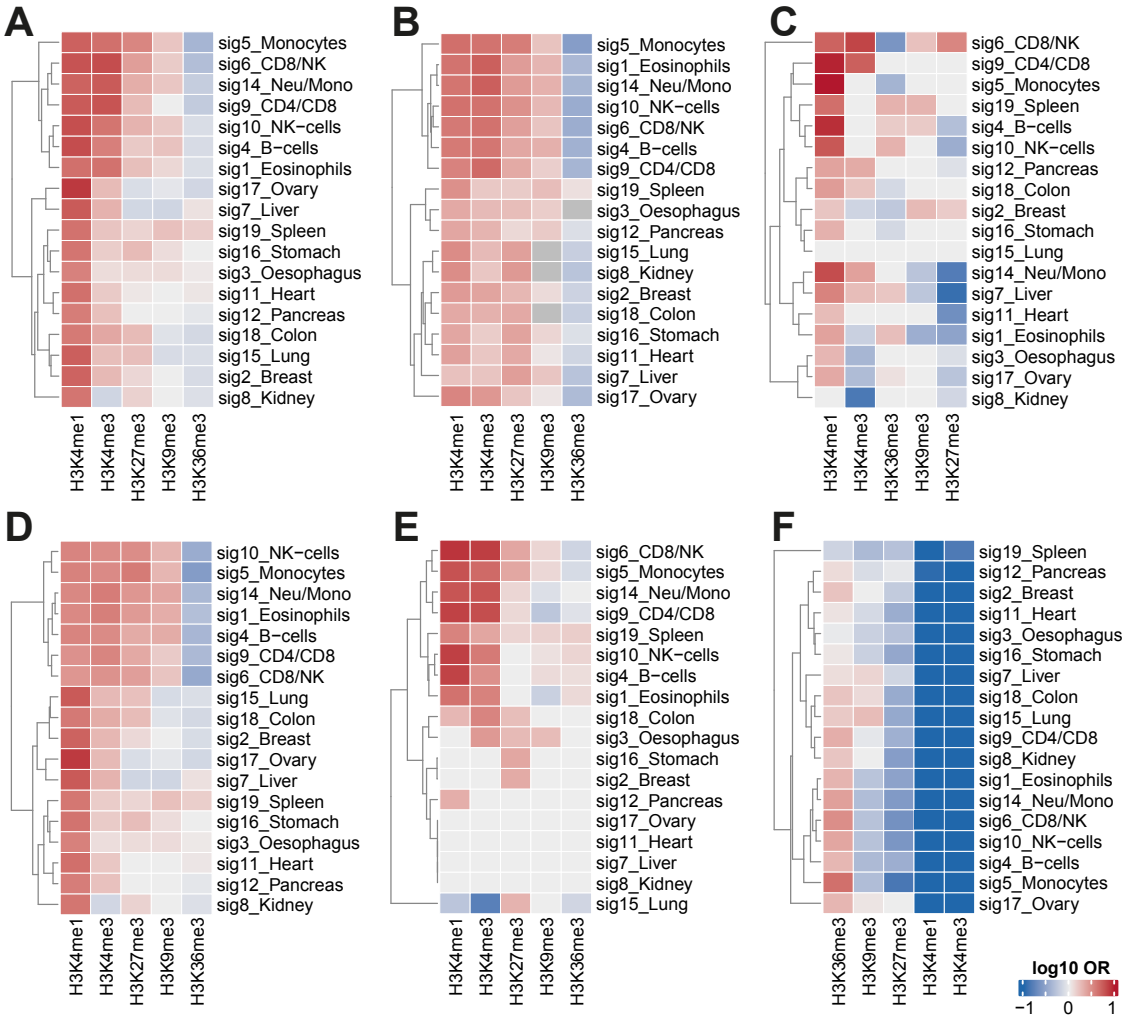


Figure 3.9: Over-representation of chromatin marks at signature-defining blocks The blocks of each signature were matched with their corresponding tissues. A) All important blocks; B) Blocks shared across more than 10 signatures; C) Blocks important for only one signature; D) Blocks important in solid tissue, but not blood signatures; E) Blocks important in blood signatures, but not solid tissue. F) The odds ratio of the blocks important for a given signature and that overlap each chromatin mark being methylated. The odds ratios were calculated using Fisher exact tests on 2x2 contingency tables. Grey squares represent p values > 0.05.

I then performed a series of Fisher exact tests on several sets of blocks. I tested how each of the marks is enriched for in all the important blocks (Figure 3.9 A).

Then, I focused on blocks shared across multiple tissues (important in more than 10 signatures), and blocks which are high-weight in one signature only (Figure 3.9 B and C). For the remaining two analyses I evaluated the differences between blocks included mostly in signatures describing solid tissues (Figure 3.9 D), and in the more blood-specific ones (Figure 3.9 E). In each of the scenarios, the presence of high-weight blocks can be explained by the H3K4me1 mark. Focusing on the distinction between single-signature and multiple-signature blocks, there is a stark difference in the representation of H3K4me3 blocks in important signatures. This mark is frequently associated with CGI promoters and regions that flank active transcription start sites. With the exception of just a few signatures, the presence of the H3K4me3 mark is negatively associated with single signature blocks and positively associated with tissue-wide blocks. Both of these, H3K4me1 and H3K4me3, are associated with lower levels of block methylation (Figure 3.9 F). Strong differences can also be observed with the H3K27me3 mark, typically associated with repressed regions and bivalent enhancers. Tissue-wide blocks for all signatures are associated with the presence of this mark; the opposite is true for tissue-specific blocks. Although H3K36me3 and H3K9me3 marks have variable enrichment within the single-cell-specific blocks, show distinct patterns in multi-tissue-specific blocks. H3K36me3 is associated with genic enhancers and transcription and is associated with elevated methylation levels, while H3K9me3 is associated with ZNF genes, repeats, and heterochromatin (Ernst and Kellis 2012; Kundaje et al. 2015).

Each of the chromatin marks above is associated with specific regions of the genome, and their occurrence is often overlapping. To gain another perspective on the regions represented by tissue-specific blocks, I performed a similar over-representation analysis on a combination of chromatin marks, known as chromatin states (Ernst and Kellis 2012). These were derived using chromHMM, which is a multivariate hidden markov model trained on the integrated epigenomes from Roadmap Epigenomics to identify 15 chromatin states associated with active and repressive functions. For each signature and its paired tissue, I calculated the number of high- and low-weight blocks that overlap each of the 15 chromatin states and performed Fisher's exact test to obtain odds ratios, divided into subsets of blocks as described above (Figure 3.10). As suggested before, the bulk of signature-defining blocks is mostly present in states associated with sites flanking active transcription (TxFlnk), enhancers (Enh) and genic enhancers (EnhG)(Figure 3.10 A). There is a fairly uniform under-representation of transcribed (Tx), weakly transcribed (TxWk) and quiescent (Quies) states. Panels 3.10 B and C suggest

the repressed Polycomb state (ReprPC) to be the least tissue-specific, as it is over-represented in blocks shared across the signature and depleted in most of the single signature-specific cases. Additionally, most panels suggest that while enhancer, genic enhancer and transcription flanking states are enriched in nearly every scenario, blood-derived signatures tend to have strong enrichment around TSS flanking sites (TssAFlnk), and active TSS (TssA).

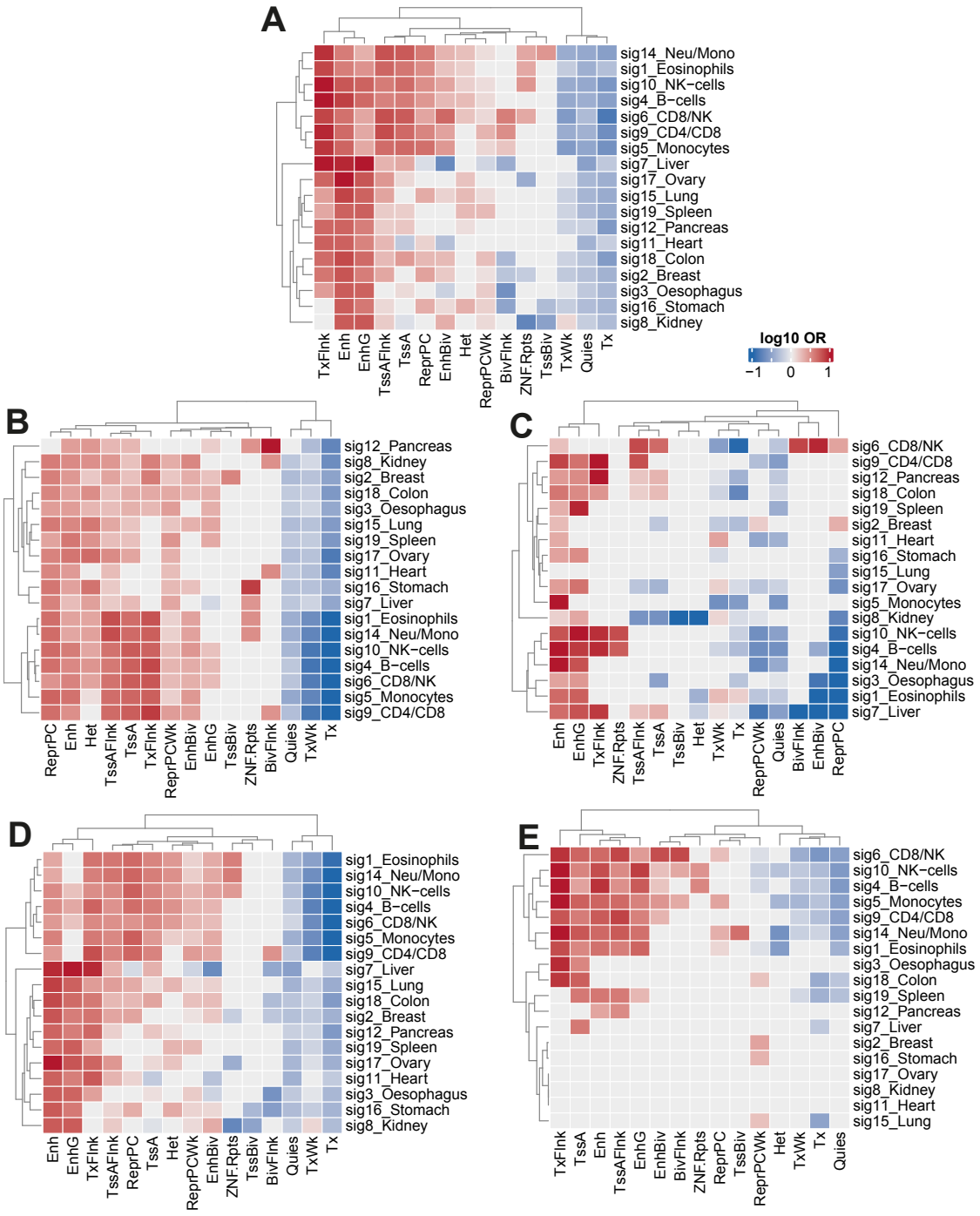


Figure 3.10: Over-representation of chromatin states at signature-defining blocks The blocks of each signature were matched with their corresponding tissues. A) All important blocks; B) Blocks shared across more than 10 signatures; C) Blocks important for only one signature; D) Blocks important in solid tissue, but not blood signatures; E) Blocks important in blood signatures, but not solid tissue. The odds ratios were calculated using Fisher exact tests on 2x2 contingency tables. Grey squares represent p values > 0.05.

3.2 Hydroxymethylation

Having identified and described the locations of tissue-specific 5mC, I investigated whether these positions are relevant only due to their lower levels of methylation or if it could be due to the presence of additional 5hmC. To do this, I took advantage of the availability of the 5hmC profiles of the samples included in the atlas.

I processed the CAPS-sequencing modification calls in a similar fashion to the 5mC, with two small, but crucial, modifications. Firstly, while 5mC is usually symmetric on both strands of the CpG, 5hmC tends to be detected on one side only. Instead of merging the CpGs from both strands into pairs, I kept the strands separate to account for the lack of symmetry if necessary in further analyses. Secondly, I removed the SNVs identified by the allele-counting method described in the previous chapter. The filtering removed mostly CpG sites with 5hmC beta value of 1 or 0.5, suggesting the presence of a homo- or heterozygous C>T change. Beta values this high are rather unexpected at non-variant positions, given that 5hmC levels usually fall between 0 and $\tilde{0}.3$. After filtering, I still found several positions with abnormally high beta values, and I found these to overlap SNPs in the dbSNP database, which most likely were missed by the allele-counting method. Although these "residual" SNPs were unlikely to impact the results from the TAPSβ atlas, such high values can skew any calculations using 5hmC. To remove potential confounders without removing too many sites in general, I only removed SNP-covered CpGs with beta values larger than 0.1, assuming that any lower 5hmC levels will have a negligible effect on the analysis. This led to the removal of additional $\tilde{2}00$ CpGs per sample.

To look into the landscape of 5hmC modifications across the highly variable blocks, I reconstructed the heatmap in Figure 3.1 by taking the average beta values per CpG block, and plotting them in the same order as defined by the original clustering (Figure 3.11). It is clear that some high-5hmC sites correspond to the high-weight blocks of CpGs with low/medium methylation values in some tissues such as ovary, liver, and B cells, and are almost completely absent in colon or T-cells. Additionally, there is a clear pan-solid-tissue section towards the top of the plot, mimicking the 5mC distributions. This suggests that in several tissues, lower, tissue-specific methylation values may in fact cooccur with 5hmC.

To further illustrate the point, Figure 3.12 shows the relation of 5hmC with the block weights. In this figure, the 5hmC peaks tend to occur in the blocks of

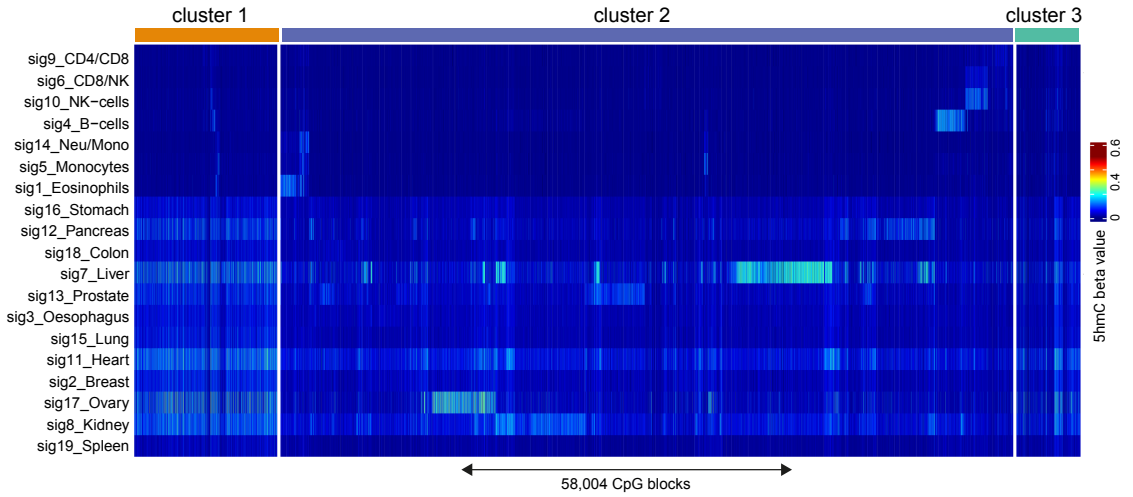


Figure 3.11: 5hmC levels at signature-defining blocks
Average 5hmC value of the blocks calculated using samples with the highest contribution of the corresponding signatures. The rows and columns are ordered according to the k-means clustering of the matrix in Figure 3.1.

the mid-weights, while still being included above the 5mC-defined cutoff point. The distribution varies across the signatures, and a clear second-peak is visible in signatures of high-5hmC tissues, such as ovary. I initially hypothesised that levels of 5hmC would be positively correlated with block weights; however, while biologically plausible, the correlation is not observed due to NMF favouring of fully hyper- and hypomethylated blocks, which by definition does not leave space for further 5hmC modifications.

Next, I have tested whether the 5hmC sites tend to be enriched in certain genomic regions. As above, knowing the mean beta value for each block, I calculated the odds ratios for a site to have a 5hmC beta value higher than 0.15 in blocks overlapping each of the previously annotated regions. The results are presented in Figure 3.13. Only a few regions present a uniform pattern across all signatures – for example, 5hmC is over-represented in introns and CGI shelves and depleted in intergenic regions, exons, or CGIs (grey squares indicate either high p-value or not enough data to calculate the coefficients). The remaining patterns vary depending on the signatures and are influenced by the number of high-weight blocks. For example, enhancers and open sea regions are depleted in several signatures on the bottom of the heatmap, while they are 5hmC-rich in the remaining signatures. If we look at Figure 3.11, we may notice that the former group tends to not have tissue-specific 5hmC signals, and the sites are dispersed throughout the heatmap. At the same time, the latter group presents more clearly tissue-specific patterns.

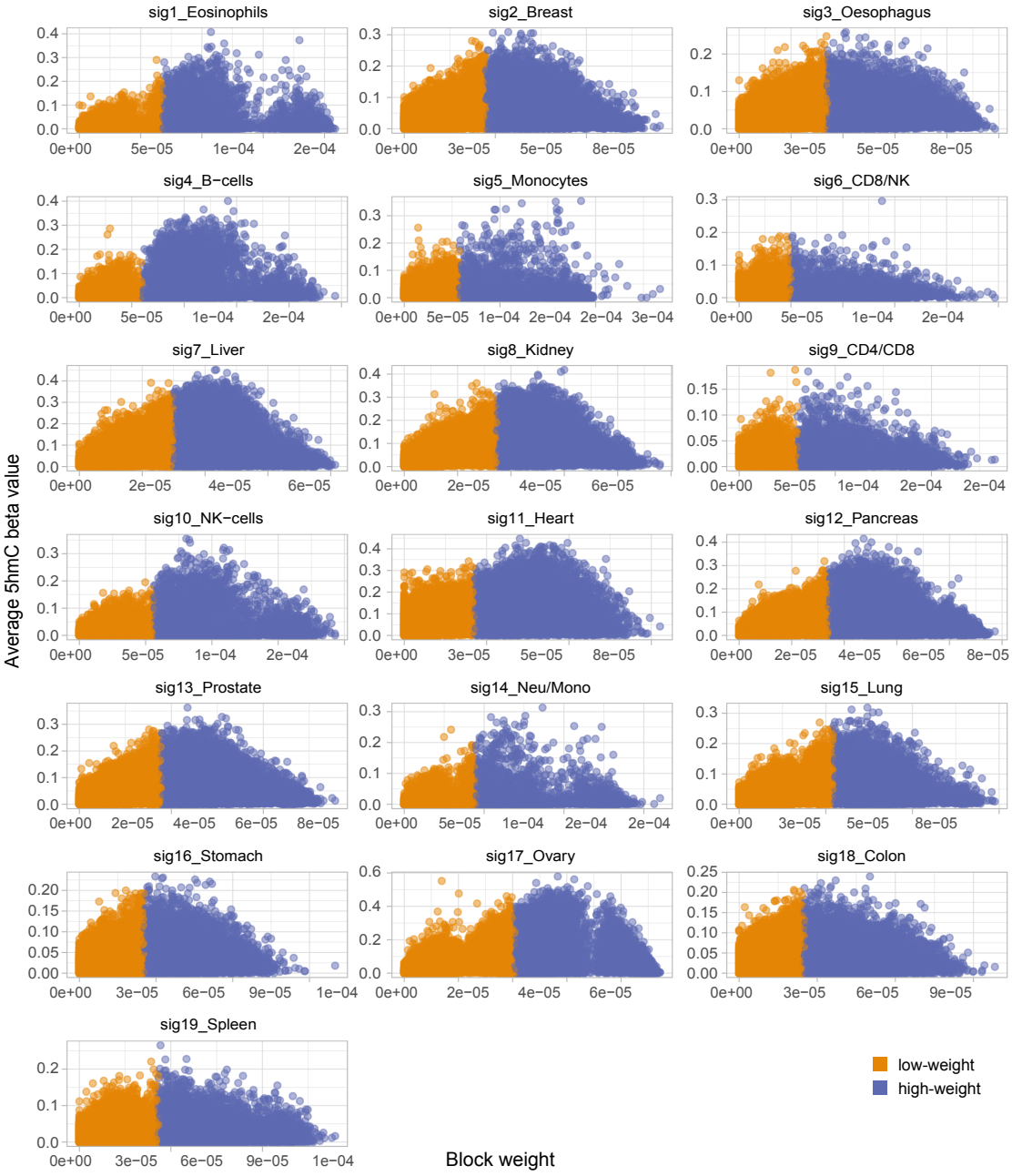


Figure 3.12: Correlation between 5hmC beta values and block weights
The average 5hmC beta values per block were calculated from samples with the highest contribution of the relevant signatures and plotted against their coefficient weight assigned by NMF.

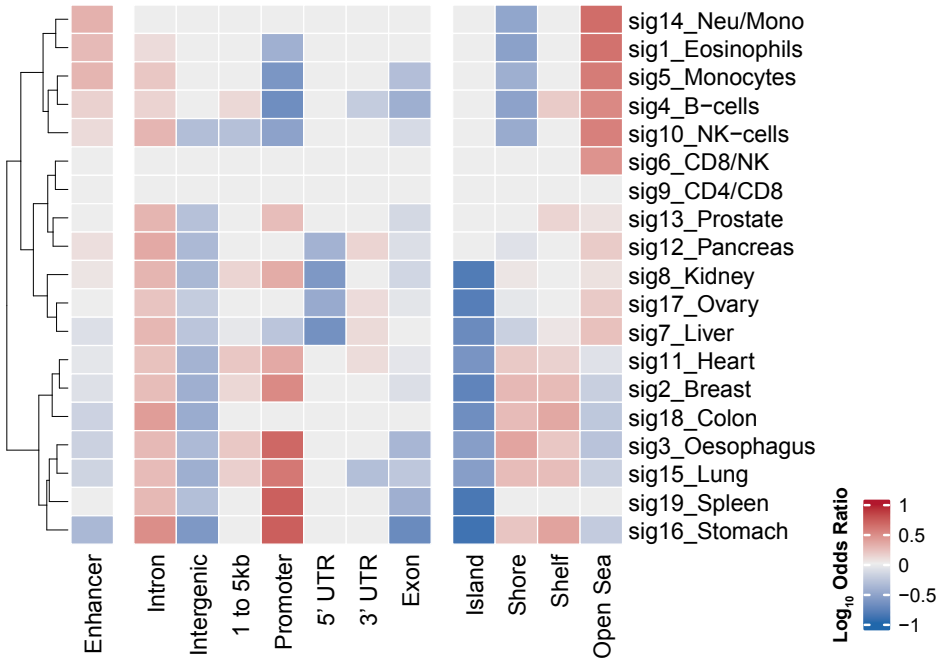


Figure 3.13: Over-representation of genomic regions in 5hmC-rich blocks
Grey squares indicate either p-value > 0.05 or not enough data to calculate the coefficients.

Another difference can be observed in the difference of 5hmC in the promoters, which is especially strong in the groups of low 5hmC specificity.

3.3 Discussion

In this chapter, I explored the nature of the cell type specificity of the previously selected blocks. The semi-unsupervised method of deconvoluting the signals from various tissues correctly identified mostly specific sites which overlap with solid tissue/cell-type-specific genes, enhancers, and chromatin states, further validating that the method is targeting real biological signals.

Firstly, I investigated the coefficient weights across all samples to visualise the "black box" of the NMF deconvolution. As expected from the input matrix, several blocks share similar weights across signatures, or are in sufficiently similar positions in the weight distribution plots to be classified as important across several signatures. The range of these similarities and overlaps varies, however 75% of the blocks are truly unique for each signature, and these have the strongest signals from the functional analyses that followed. I decided to keep the two nonspecific clusters in the matrix to preserve the semi-supervised nature of the analysis. It must be remembered that Figure 3.1 illustrates only the important blocks for

the clarity of the visual interpretation, and in reality the entire matrix has more lighter spots corresponding to CpGs with lower weighs which did not make it past the cutoff. The spectrum of weights is important to keep in mind, especially when interpreting the blood-cluster results – while it may be confounding the functional analyses by representing nearly all signatures, in reality, it is the blood signatures which have the highest coefficients in that cluster, and it is likely that during deconvolution these sites have neglible effects on the fit. These issues of overlapping important signals require a wider discussion, which I will undertake in the last chapter.

The majority of the signature-defining blocks proved to be biologically relevant and stood the analysis from several angles. The analysis that provided the highest granularity was the gene-based ORA, which was based on sets of tissue-specific genes obtained from the Human Protein Atlas. Combining these results with indirect enhancer associations further improved the specificity and a clear signal was seen using a separate enhancer-based database, not taking genes into account. The identity of the non-specific clusters proved challenging to analyse, both using Gene Ontology gene sets, and the tissue-specific ones. However, it is worth noting, that in the gene-based ORA including all clusters together did not drastically change the obtained results, further proving that the addition of the extra clusters is not likely to undermine the biological signals.

It is also evident that the selected blocks have strong association with certain chromatin marks, states, and the general measure of chromatin accessibility. This confirms that the higher-order DNA structure is closely tied to tissue-specific methylation. The ChIP-seq ORA using all blocks are consistent with previous findings, such as the specificity of the H3K4me1 mark, the 5mC status associated with each modification, and the most tissue-specific chromatin states as defined by ChromHMM (Kundaje et al. 2015; Ernst and Kellis 2012; Loyfer et al. 2023). Interesting points arise from the distinctive patterns between the signatures of blood and solid tissue, revealed by a particularly strong enrichment of histone marks and chromatin states associated with transcription start sites and promoters. As indicated in the previous chapter, both sets of signatures have slightly different natures of the blocks, and this result may correspond to the earlier observation that the blood-derived signatures are richer in hypermethylated blocks present in CGIs and promoters. This is a further indication that the blood signatures may be defined by hypermethylation of CGI promoters (and therefore their silencing). Due to the generally lower number of high-weight blocks in blood-derived signatures,

it is plausible that my method strongly favoured the CGIs because of their extreme methylation values, which are distinct from the mid-modification levels associated with enhancers, as observed both in my data and in literature (Kundaje et al. 2015; He et al. 2021).

Incorporating 5hmC signals also allows to glimpse into the reason for the informative state of the selected blocks. As described in the literature, 5hmC patterns are highly tissue-specific, sometimes to a higher degree than 5mC (Cui et al. 2020; He et al. 2021; Nestor et al. 2012). Although I did not take the 5hmC levels into account while preparing the atlas, using TAPSβ instead of WGBS resulted in the 5hmC-abundant CpG sites showing lower levels of 5mC modifications, increasing their likelihood of being included in the atlas. It is plausible that several of these sites would be omitted in the creation of bisulfite-based atlases limited to a small number of CpGs or CpG blocks per tissue of interest. Furthermore, this again reflected the difference between the identification of blood-derived and solid tissue signatures, as illustrated by lower levels of 5hmC in blood signatures, likely associated with under-representation of enhancers.

In summary, the functional analysis confirmed the biological relevance of the selected CpG blocks, and shed light into the tissue-specific methylation patterns, suggesting that limiting the CpG site choice to most hypo and hypermethylated sites as done in other methylation atlases may exclude a lot of informative sites, especially those covering enhancers.

4

Applications in Oesophageal Adenocarcinoma

Contents

The molecular analysis of tumour biopsies is a challenging task. Tumours are heterogeneous and biopsies are imperfect; as a result, bulk sequencing of such samples tends to represent a mixed mixture of a variety of cell types, rather than the "true" tumour signal. This does not necessarily reflect an "impurity" of the sample, as tumours rarely exist on their own, and the immune microenvironment present within the tumour has been described to have an impact on tumour progression and therapy outcomes (reviewed in Fridman et al. 2012; Jin and Jin 2020). The methylation profiles obtained from bulk sample data are representative of the tumour and the context the tumour is in. Knowing the methylation profiles of a variety of normal cell types, one can use the data to decipher the contents of the bulk tumours by accessing the similarity of the bulk methylation landscapes to the reference, finding an optimal combination to reflect the tissue.

In this chapter, I endeavour to use the tissue-specific CpG blocks to deconvolute and investigate oesophageal adenocarcinoma (OAC) samples obtained from the

LUD2015-005 clinical trial. I apply the method to find the underlying similarities between the samples, compare the differences between the tumours obtained at different timepoints, and attempt to find signals linking the methylation profiles to clinical outcomes.

4.1 OAC and LUD2015-005 trial

Oesophageal adenocarcinoma is a subtype of oesophageal cancer characterised by malignant epithelial cells originating from glandular tissue (Smyth et al. 2017). Its prevalence has seen a marked increase in recent decades, especially in Western countries (Fitzgerald 2004; Devesa et al. 1998; Bray et al. 2018). Despite advances in early detection and treatment, the prognosis remains poor, with high mortality rates; many patients present in advanced stages when therapeutic options are limited (Pennathur et al. 2013). For these inoperable patients, treatments usually consist of fluoropyrimidine and platinum chemotherapy (CTX), but both result in a median overall survival (OS) of less than one year. (Cunningham et al. 2008; Jatoi et al. 2006). In 2021, the FDA approved a regime combining immune checkpoint inhibitor (ICI) and CTX (ICI + CTX) for inoperable gastro-oesophageal patients (FDA 2021a; FDA 2021b). Although this form of immunotherapy led to substantial improvements in outcomes in patients suffering from various cancers, long-term benefits are limited to a small fraction of treated patients (Robert 2020). These daunting statistics motivated a wide range of research dedicated to the search for biomarkers associated with long-term benefit from ICI-based therapies. The studies identified a variety of key predictors of ICI response, including tumour mutational burden, expression of targeted checkpoint molecules, or markers of T cell inflammation (Litchfield et al. 2021). These predictors vary across cancer types and can be influenced by the addition of CTX, which can damage both healthy and cancer cells. To date, there is no successful biomarker of clinical benefit of ICI + CTX regimes in patients with OAC. The phase I/II LUD2015-005 trial aimed to bridge that gap by treating patients with inoperable OAC (and three patients with oesophageal squamous carcinoma, OSCC) with ICI alone for four weeks, followed by ICI + CTX treatment. Biopsies were taken before treatment (ScrBsl), after the ICI window (Immonly), and after treatment with ICI-CTX (C6D22) (Figure 4.1). Complex molecular profiling was performed on all biopsies. The analysis of WGS, bulk RNA sequencing, and single-cell RNA sequencing was recently published by Carroll et al. 2023, who identified high tumour monocyte content and mutational burden as good predictors of ICI+CTX outcomes. Whole genome methylation profiles from 24 patients were also obtained using TAPS with the aim of identifying additional methylation markers that predict the response to treatment.

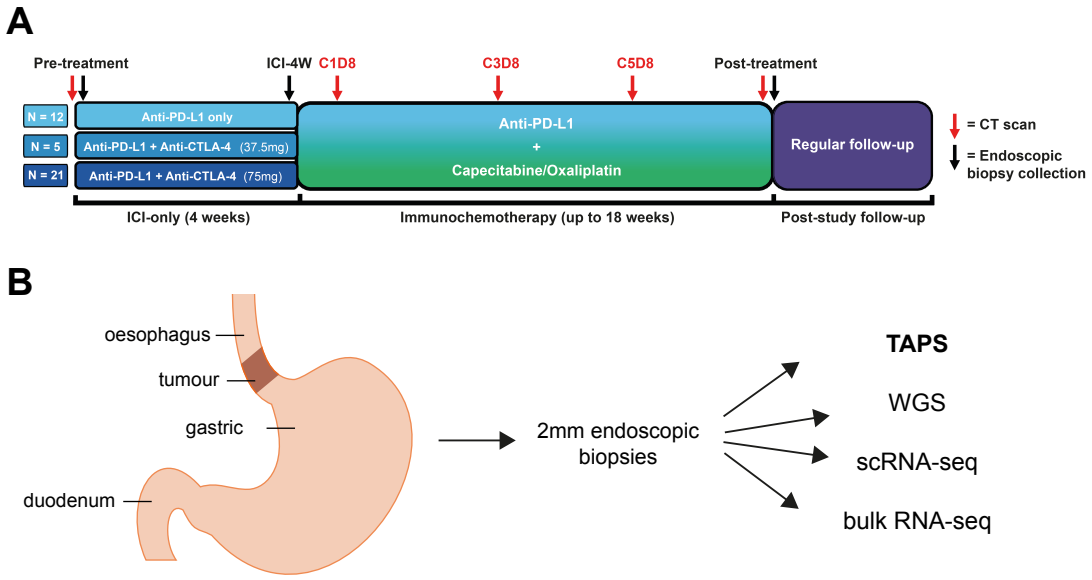


Figure 4.1: Overview of the LUD2015-005 clinical trial
A. ICI-only and ICI+CTX treatment regimen and endoscopic biopsy collection time points. Panel A taken from (Carroll et al. 2023). B. Schematic of biopsy collection sites and molecular profiling performed on the samples.

4.2 Methylation profiles of OAC samples

To describe the epigenomic characteristics of the trial samples, I first calculated the average methylation values of the 58,004 blocks identified in the first chapter. For the initial visual inspection of the samples, I plotted the highly variable blocks as a heatmap (Figure 4.2). As can be seen, there are tumour samples clustered together with each of the included healthy tissues, as well as visibly more distinct clusters of samples with tumour-specific methylation profiles. There are a few unexpected mismatches, such as one seemingly oesophageal sample (071-014_ScrBsl_oesophagus) showing a more tumour- than oesophageal-like methylation pattern. Samples from the same donors tend to present similar methylation profiles, although they are separated in multiple instances.

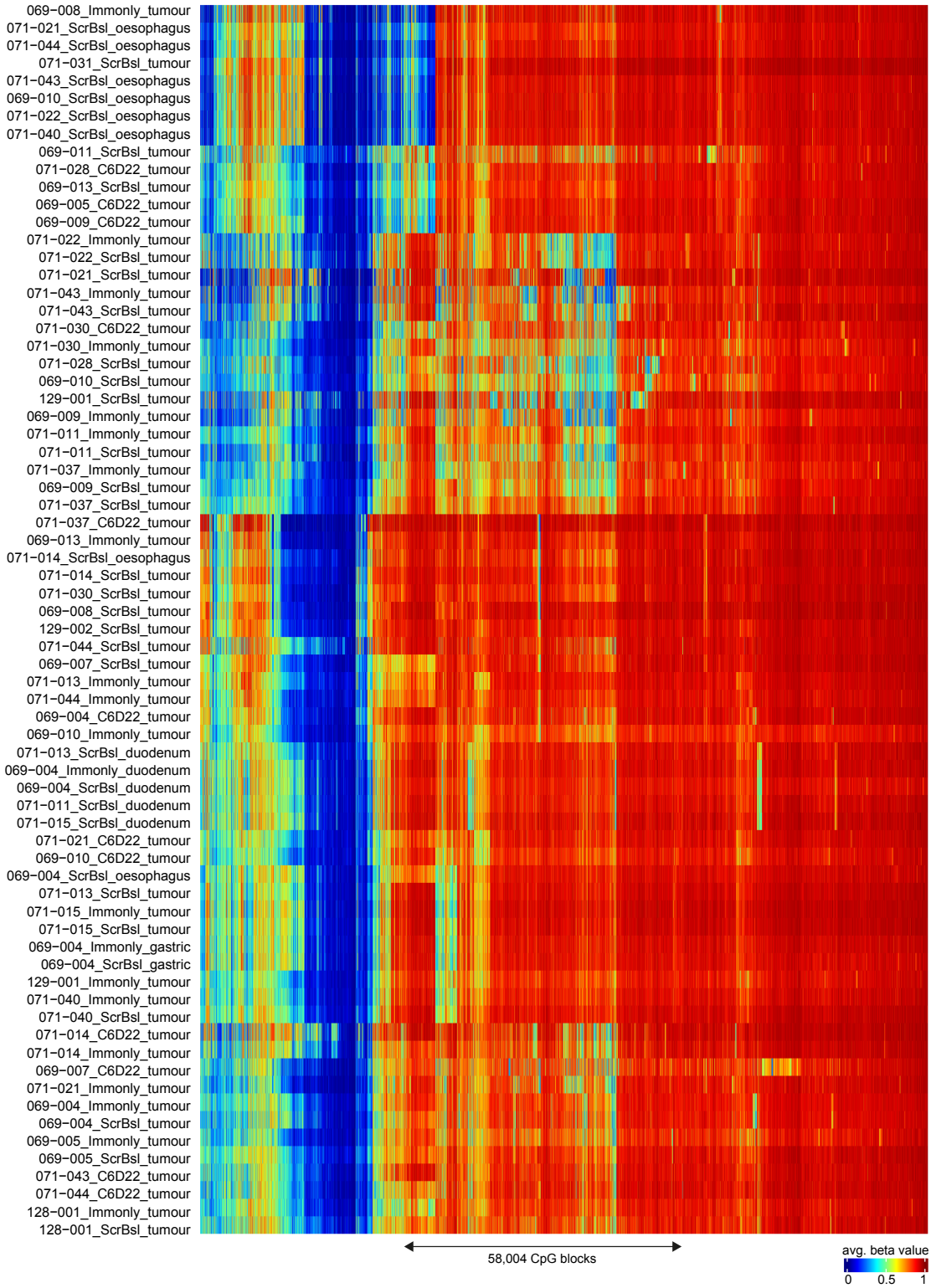


Figure 4.2: Methylation profiles of the trial samples using the TAPSβ atlas blocks The average beta value per CpG block. Rows and columns were clustered using the Ward's method (dendrogram not shown).

To gain more insight into the sources of differences in these methylation patterns and decipher the cellular components of the samples, I used NNLS to fit the trial samples to the coefficients obtained using the reference TAPSβ atlas. (Figure 4.3). Looking at the contribution heatmap, one can see clusters similar to those of the previous heatmap, with some samples clearly following healthy tissue patterns. I included the healthy tissues into the deconvolution as positive controls to measure how well they are deconvoluted to their correct signature counterparts. Healthy duodenum samples do not have their own signature, but are split between the signatures of the stomach and colon. The unassigned tumours fall into four additional groups - first, there is a subset with high contribution of the neutrophil / monocyte signature, next is an oesophageal-like group with a substantial contribution from the stomach and colon signatures, tumours with stomach signature but strong admixture of colon and immune cell type signals, and, among those, several samples without a clear signal from any source. Interestingly, the only OSCC sample in the dataset, 069-011, had a strong oesophageal signal with minimal stomach signature admixture.

When investigating the residual values after fitting each sample, one can notice large differences between the samples. The residual values tend to be lower for healthy tissues (with the exception of the oesophagus) and higher in tumours. I compared the norm of residuals against the estimates of tumour DNA content from several samples and found a strong correlation between the two (Figure 4.4). This may suggest that the more challenging it was for NNLS to fit the data correctly, the more tumour there is in the sample. This could reflect the fact that there is no pure tumour reference in the input matrix. If we assume that the tumour has a very distinct methylation profile, which would not be picked up by the NNLS and hence result in a high norm of residuals, then we may assume the remaining signal of contributions as coming from the tumour microenvironment, neighbouring tissues, or from more healthy tissue resembling samples. This conclusion can be taken further to potentially scale the contributions of signatures in the samples to reflect the amount of tumour, and make the between-sample comparisons more fair.

Knowing the good correlation between purity and residuals, I constructed a simple linear model to predict the tumour DNA content of samples that did not have these values described experimentally.

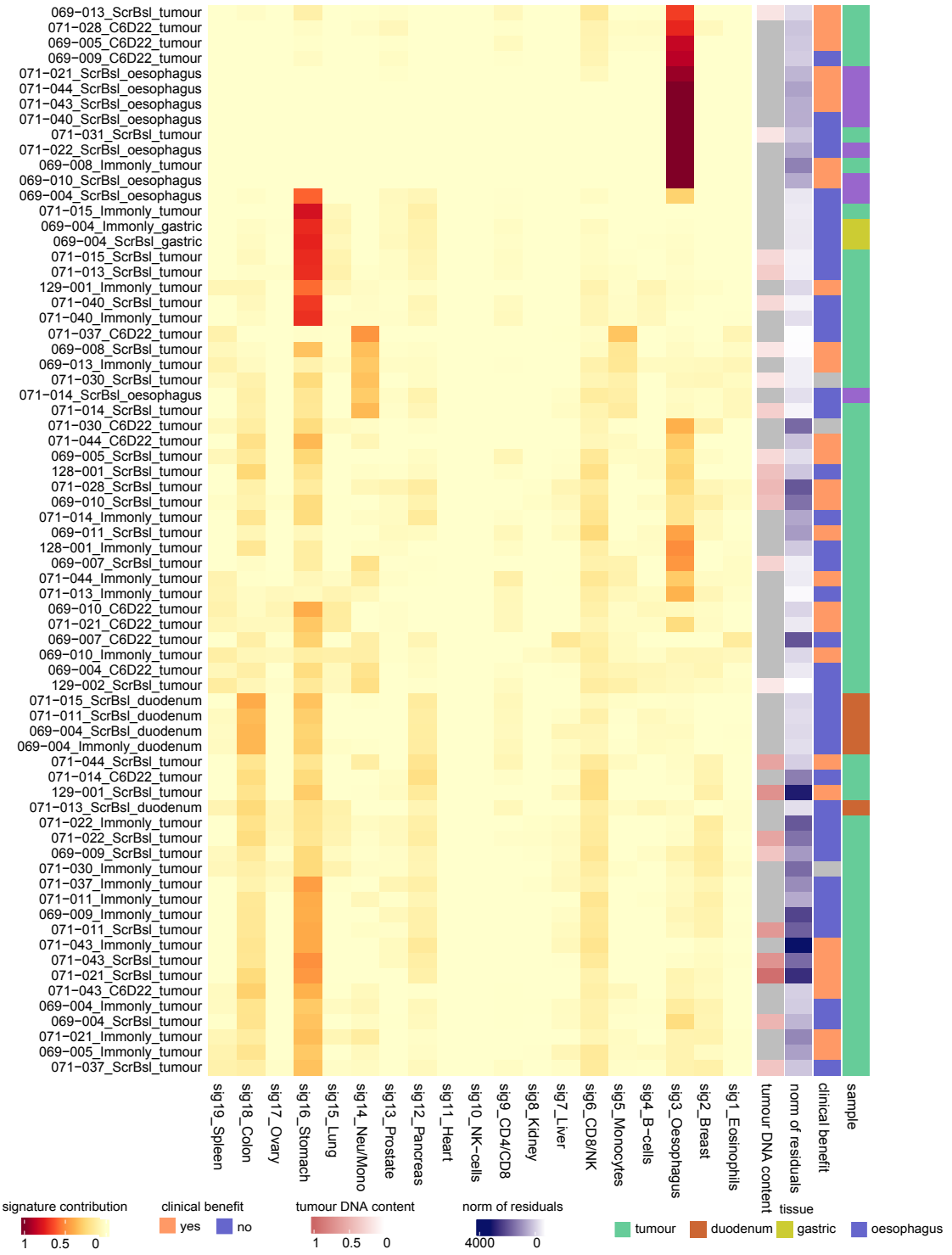


Figure 4.3: OAC trial signature contribution heatmap
Contributions of each of the 19 TAPSβ atlas signatures obtained using NNLS. Total contributions for each sample were scaled to 1 to allow comparison between samples. The columns were clustered using hierarchical clustering. Tumour DNA content annotation is included in samples for which the predetermined values were available. Norm of residuals was obtained directly from the NNLS fitting. Clinical benefit was defined as attaining 12 months of progression-free survival. ScrBsl, Screening baseline; Immonly, Immunotherapy only; C6D22, Immunotherapy + chemotherapy.

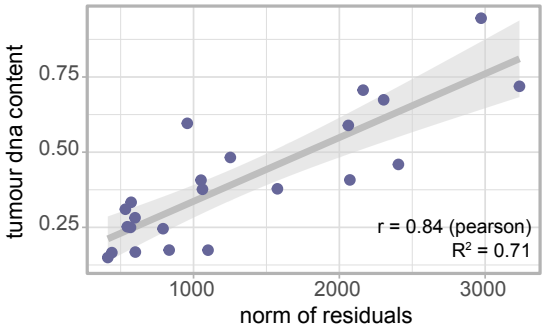


Figure 4.4: Correlation of tumour DNA content and NNLS norm of residuals
Tumour DNA content is strongly correlated with the norm of residuals obtained from NNLS fitting. The grey line and ribbon illustrate the linear model fit and 95% confidence intervals. The correlation was calculated using Pearson's method (p value < 0.001).

4.3 Signature contributions as biomarkers

Calculating the contributions of methylation signatures in the trial samples provided a range of measurements that could be tested for potential biomarker properties. Knowing the clinical outcome and overall survival of the patients allowed me to perform a series of survival analyses to detect any signal that could potentially predict the outcome of treatment. Figure 4.5 shows a rearranged version of the previous heatmap, split by patients, ordered by timepoints and reflecting the estimated tumour DNA content as an annotation, providing a visual reference for the analyses described below.

Firstly, I computed the survival models using the contribution of signatures at the ScrBsl timepoint, dividing them into high- and low-contribution groups (with respect to the median of contributions in the given signature). I performed this analysis using four methods to calculate the signature abundance: contribution scaled to 1 (values shown in the heatmaps above), contribution scaled to 1 corrected for tumour DNA content, contribution as it appears in the NNLS output (not scaled to 1), and last, not scaled contribution corrected for tumour DNA content (details are described in Methods 6.6). This allowed for the exploration of all cases, allowing me to notice potential biases introduced by the different value corrections. Survival analyses using all signature contributions revealed no significant results in any of the cases. In several instances, the differences were minimal, especially for signatures with very low exposure in any of the samples. Interestingly, the high DNA content of the tumour appeared to be slightly favourable in terms of overall survival, although the test did not produce a significant result ($p = 0.1$).

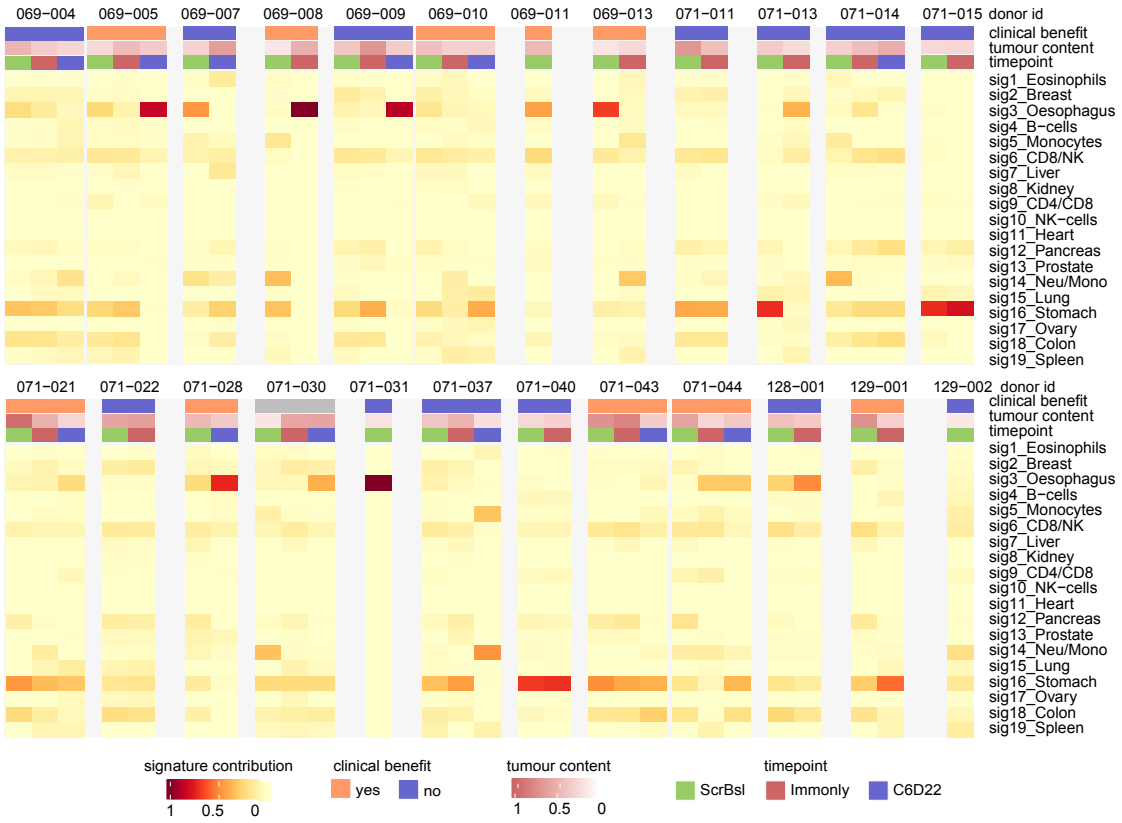


Figure 4.5: Patient-focused OAC trial signature contribution heatmap
The heatmap illustrates the same values as in Figure 4.3, with columns grouped by patient and ordered based on the sample collection timepoint. Tumour DNA content annotation reflects the values from the previous plot, with missing data predicted from the residual-based linear model. Clinical benefit was defined as attaining 12 months of progression-free survival. ScrBsl, Screening baseline; Immonly, Immunotherapy only; C6D22, Immunotherapy + chemotherapy.

Due to the availability of data from multiple time points from the same patients, I also investigated fluctuations in signature contributions throughout treatment. I calculated the differences between each of the four values described before in three scenarios: ScrBsl/Immonly, ScrBsl/C6D22, Immonly/C6D22, and divided the samples accounting for the decrease or increase in signature contribution.

Interestingly, in this case two significant hits were reported in the difference between ScrBsl and Immonly timepoints. One included the stomach-specific signature, which is widely present in almost all the samples analysed (Figure 4.6 A). According to the survival analysis, a decrease in the contribution of the stomach signature between these time points is associated with a lower probability of survival. This may be a reflection of the progress of treatment in which stomach-like tumours are removed. Additionally, the kidney-specific signature was

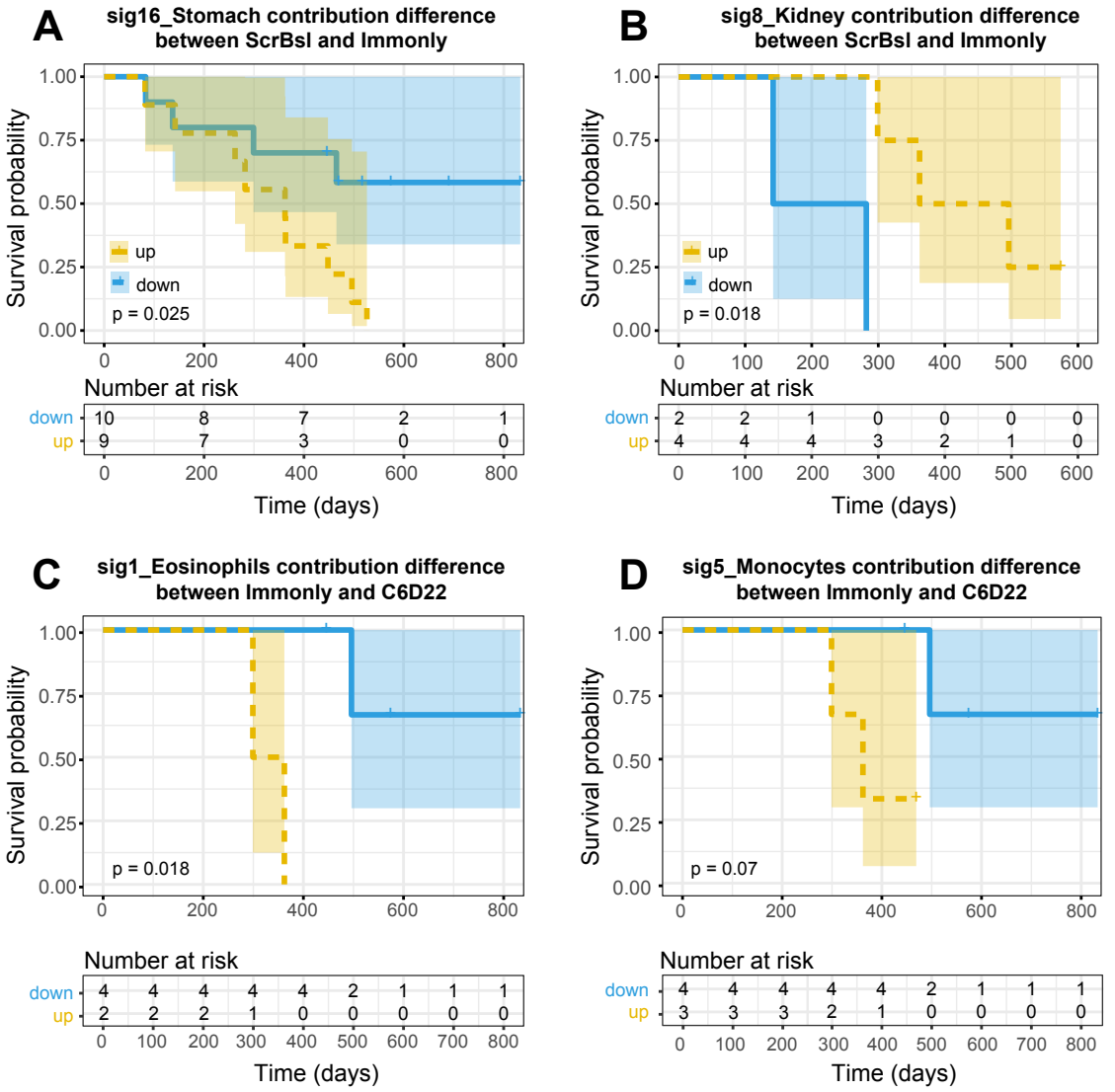


Figure 4.6: Survival analysis of contribution differences across treatment timepoints Kaplan-Meier plots showing the survival probability according to changes in four signature contributions measured at the ScrBsl/Immonly (A, B) and Immonly/C6D22 timepoints (C, D).

associated with survival, where the samples with an increase of the signature had higher survival probability (Figure 4.6 B). This is more challenging to interpret, especially given the low sample size in this analysis caused by the marginal contribution values of this signature. Given this and the unlikely match that is the kidney and oesophageal cancer, it is quite likely a false positive caused by a low sample size.

I compared the contributions at the Immonly and C6D22 time points in the same manner, and this analysis indicated that an increase of two signatures had a

negative impact on overall survival: signature 1 and 5, associated with eosinophils and monocytes, respectively (Figure 4.6 C, D). As above, the low sample size creates a challenge with the interpretation of the results. Eosinophils have been reported to predict outcome differently in a variety of tumours. On the other hand, although the monocyte results were not quite significant, I included the result because it indicates an opposite effect to what was observed in RNA-based deconvolution, where high monocyte content was associated with better clinical outcomes (Carroll et al. 2023).

4.4 Discussion

In this chapter, I showed the applicability of the tissue-specific methylation atlas in a cancer research setting. Fitting the 58,004 CpG blocks from the OAC trial samples to the NMF-derived coefficients revealed groups of tumours of distinct methylation profiles. These groups fell into three categories - oesophageal-like, stomach-like and unassigned, where there was no dominant signature. The patterns of the oesophageal and gastric signals are possibly related to the common discussion in the field about the cellular origins of OAC and its common precursor Barrett's Oesophagus (BO), a pre-malignant lesion that forms as a response to gastro-oesophageal reflux (Peters et al. 2019; Siewert and Ott 2007). In addition to difficult demarcation at the gastroesophageal junction (GEJ), oesophageal and gastric adenocarcinomas show striking molecular similarities, prompting debates on the classification of the conditions and the utility of histological distinctions (Suh et al. 2012; Leers et al. 2009; Siewert and Ott 2007). In fact, they were found to be more similar to each other than OAC is to OSCC (Kim et al. 2017) – which could explain the very low stomach signature contribution in sample 069-011, which represents the only TAPS-sequenced OSCC sample in the trial. The origins of BO are also subject to fierce debate, with numerous evidence suggesting its emergence from proximal gastric cells (Quante et al. 2012; Wang et al. 2011). This evidence has been widely challenged due to the reliance on mouse models, but emerging human studies point to similar conclusions (Nowicki-Osuch et al. 2021).

Because of intricate analyses and single cell work required by the nature of these "origin-finding" studies, it is very unlikely that my deconvolution can contribute to the topic, and this was not its purpose. However, these uncertainties can provide an explanation of the so widely present stomach signature in nearly all of the samples, possibly capturing the "gastric-ness" of the biopsied tumours. It

is challenging to interpret the meaning of the relative differences between the signature contributions within and between the samples. No biopsy and no tumour are the same, and simple variability in the exact site of biopsy collection could lead to stronger oesophageal or gastric signals, depending on the distance from the GEJ.

The correlation of the norm of residuals with the tumour DNA content was an unexpected but logical finding. Calculating the tumour purity and ploidy from bulk samples is a challenge, and a variety of methods are available. These range from microscopic evaluation by pathologists to tracking a specific mutation, to estimation based on copy number changes. The latter method is frequently used in NGS studies, sometimes coupled with histology reports. An example software for this purpose is ASCAT, which bases the calculations on SNP counts (Carter et al. 2012; Raine et al. 2016). In some cases, however, when such estimates are unavailable or suffer from low goodness-of-fit measure, one can benefit from additional measures to obtain or confirm the estimates. The strong correlation between the norm of residuals and tumour DNA content may be such measure. Because it relies on the badness of fit of the reference atlas, it may be influenced by changes in the reference or the analysed tumour. An atlas with wider array of tissues, possibly even reflecting tumours, could lead to lower residual values changing the end result. From a few analyses which I have not included in the thesis, I can say that the strong correlation remains despite applying a larger atlas, such as from tissues obtained by Loyfer and colleagues. The norm of residuals decreases drastically but proportionally, not changing the final shape of the plot. I have also applied this pipeline to a separate ovarian cancer dataset from Gull et al. 2022, which showed a very close similarity (results not shown).

I hypothesised that the norm of residuals or tumour DNA content estimate could be treated as a fraction of tumour that did not fit into the reference atlas samples. To test that, I corrected the contribution values to illustrate the predicted tumour content and reduce the contribution of non-tumour signal accordingly. If these assumptions were true, one would expect the following survival analyses to favour calculations adjusted for tumour DNA content, assuming that in nonadjusted samples the contributions would be disproportionally inflated resulting in noise. The survival analyses, however, suggested the opposite, and the only significant hits came from unadjusted contributions. On one hand, this suggests that scaling contributions to one is an appropriate measure to enable comparability of the results. On the other hand, it indicates that correction of the NNLS fit results with

the tumour content may not be effective, and one must analyse the data under the assumption that the NNLS fit is capturing the entirety of the sample methylation pattern, with the residuals not being a form of unassigned methylation values specific to the tumour.

Survival analyses suggested that changes in the contribution of certain signatures across timepoints were associated with better overall survival. While these observations would not classify as useful patient stratification biomarkers, as levels of none of the contributions before treatment were associated with survival, they could provide interesting insight into the changing tumour environment and the progression of therapy. For example, the favourable decrease in stomach signal (Figure 4.6 A) may reflect the gastric-like properties of OAC, and the decrease in their prevalence as a marker of successful treatment. The remaining results suffer from much smaller sample sizes, caused by both a smaller number of TAPS-sequenced longitudinal samples and in some cases the lack of detected signature signal. Interestingly, eosinophils play important role in oesophageal pathologies. Eosinophilic oesophagitis (EoE) is a chronic inflammatory condition. Although to date no association has been found between EoE and OAC, EoE is believed to have protective effects against OSCC (Syed et al. 2017). Furthermore, high levels of eosinophils have been linked to favourable outcomes in patients with OSCC (Abe et al. 2011; Dos Santos Cunha et al. 2023; Ohashi et al. 2000; Ishibashi et al. 2006). This may potentially suggest some role of eosinophiles in OAC progression; however, this finding is very limited by low total contributions and sample size. With respect to monocytes, I chose to illustrate the result despite not meeting the significance cutoff. Carroll et al. 2023 and colleagues performed RNA-seq-based deconvolution of ScrBsl tumours from trial patients and found that increased monocyte content is associated with higher chances of clinical benefit. While I did not observe such result in the methylation deconvolution, the fact that the increase in monocyte content after immunotherapy was unfavourable may suggest that successful therapy would bring this cell type's count down, rather than up. A larger sample size and refined pipelines are necessary to investigate this relation further.

5
Discussion

In this thesis, I presented a novel method for the identification of highly-variable, tissue-specific CpG blocks across the genome. The method is based on a TAPSβ atlas representing a collection of 20 different cells and tissues. I applied the method to both heterogeneous data, and FACS-sorted cells, obtaining satisfying results in cross-validation, finding that the method required minimal optimisation to be applicable to new datasets. The selected blocks cover a wide range of genomic contexts. In the second chapter, I investigated why the selected blocks are likely to have tissue-specific properties by performing functional analyses and introducing higher-order epigenetic modification data. In particular, I showed the gene-related tissue specificity of the majority of identified blocks, followed by weaker, but significant enhancer-based specificity. The functional analysis revealed that the selected signature-defining blocks have a strong association with the H3K4me1 histone marks, and active chromatin states. The functional analysis revealed differences in the blocks obtained from blood and solid tissues. I have also showed the substantial enrichment of 5hmC at the highly-variable blocks present in most tissues. In the last chapter, I presented the practical use case of the developed deconvolution method by applying the reference 5mC atlas to samples from patients with OAC. The fit gave insight in the cellular composition of patient samples, tumour purity and found that changes in the contribution of several signatures across the time points were associated with different survival probabilities. Here, I will discuss the findings in a wider context than at the end of each section, combining the insights from all chapters.

Good-quality methylation data are the basis for creating a successful 5mC reference atlas. Quality here can be understood in several ways, beginning with correct handling of tissue material, the universal processing of sequencing data, and appropriate coverage. Until recently, publicly available whole genome datasets were limited to a handful of samples from the Roadmap Epigenomics and ENCODE projects (Kundaje et al. 2015; ENCODE Project Consortium et al. 2007), and because of this, nearly all cell type deconvolution methods were based on Illumina HM450 arrays (Houseman et al. 2016; Titus et al. 2017; Moss et al. 2018; Chakravarthy et al. 2018; Koestler et al. 2016). For the majority of the project and method development, I used the Roadmap/ENCODE datasets as the reference, as it was only January 2023 when the in-house TAPSβ atlas was sequenced, which coincided with the publication of another large WGBS-based atlas from FACS-sorted cells (Loyfer et al. 2023). The availability and quality of the public datasets shaped the method development process, as the datasets were heterogenous and varied between samples from the tissue types. Loyfer et al. 2023 applied their deconvolution algorithm to the public samples to find the same conclusion. The results showed that a large proportion of samples are composed of multiple tissues, the most extreme examples like the colon show up to 60% of its composition to be fibroblasts and macrophages. This was the reason for semi-supervised approach to the block selection, for example at the stage of genome segmentation. Assuming perfect and homogenous datasets, one would not need to merge samples from one tissue type to obtain blocks. Unfortunately, because of heterogeneity of the non-sorted samples, this step was necessary to retain enough blocks and allow for some degree of flexibility between the samples. In the future, when bigger methylation sequencing collections are available, one can perform reference-free deconvolution on each separate tissue type to identify the consensus sequences which are unaffected by inter-sample variation.

This heterogeneity presents multiple challenges. Most of the methylation sequencing data will be obtained from bulk, non-sorted solid tissues, as in the TAPS atlas. Cell sorting is labour intensive, requires higher amounts of tissue material, and sometimes the markers to identify all relevant cell types are not available. From sample to sample, heterogeneity may be introduced due to differences in biopsy sampling, potentially introducing a distinct proportion of cell types which will make up the bulk tissue. The samples may or may not come from the same developmental lineage, as multiple organs are formed from a merge of cell types that come from different layers during development. In contrast, cells obtained from peripheral blood are much easier to sort and their methylation signatures

are more clear, usually uncontaminated by different tissue types (depending on the sorting method). An example of this challenge was observed at the stage of variance filtering, when blood samples had to be filtered separately to obtain a higher number of blocks of interest, as their variance was overshadowed by the bulk tissue data. Despite the issues, bulk tissue heterogeneity may not be a "confounder"; in fact, it appreciates how complex these tissues are. By definition, sorting and filtering cells will remove certain cells from the tissue. As much as I know about the samples, I am currently unable to isolate and sequence every cell type. This results in the removal of some cells, and the methylation patterns obtained represent an incomplete landscape of the sample. Bulk sequencing, in contrast, will contain many cells, and the total methylation landscape does reflect the various cell types present in the cells. Of course, this may lead to noise and multiple tissues sharing the same blocks, but the noise can be removed or accounted for in further analyses. Depending on the potential application, both approaches have reasonable use cases. Most of the deconvolution-focused research published to date has clearly strong interest in finding sets of prognostic or diagnostic biomarkers to use in cancer and other conditions (Titus et al. 2017; Loyfer et al. 2023; Moss et al. 2018; Chakravarthy et al. 2018; Wiencke et al. 2017). For that, a small number of highly-informative sites is more beneficial than having any shared sites, as it improves the specificity. I did not limit the number of used sites to allow for wider biological exploration, as I was not constrained by the applicability in cancer diagnostics.

Whole-genome methylation sequencing is dominated by WGBS, but the collection of TAPS methods is gaining popularity due to the removal of harsh bisulfite treatment, especially relevant in the context of cfDNA. The first and last chapters showed that in terms of the atlas creation and deconvolution, the method can be used on any type of methylation data, as long as one is not looking for the smallest 5mC level changes or attempts to adhere to the strict rules of clinical quality. Using TAPSβ as the input for CpG site selection increased the chances of detecting the 5hmC signal, as at the sites of modification WGBS sequencing would show elevated levels of methylation, which could in fact be hydroxymethylation. The removal of brain tissues from the atlas was dictated by the significantly distinct 5hmC profile and low levels of 5mC, as illustrated in the heatmap and on a whole-genome scale in the appendix (Figure B.2). Substantial modifications to the pipeline would be required to incorporate brain tissues into the model, and in case of cfDNA-based biomarker discovery a few selected DMRs would work well. For biological exploration of the tissue-unique signal, I would propose to study

brain tissues separately, or tailor the analysis to identify brain-specific sites. The brain itself is made up of multiple, often poorly understood cell types and with the distinct 5hmC profiles it becomes a whole different and fascinating area of research (Jin et al. 2011b).

In addition to the heterogeneity challenges, methylated CpG sites are at high risk of C>T mutations, and the majority of mutations and SNPs in the human genome occur there. Depending on the sequencing method used, the mutations can be interpreted as unmodified (WGBS) or modified (TAPS, CAPS) cytosines. In the 5mC context, despite the much great numbers of reported modified cytosines, C>T variation may influence the selection of variable sites. I have experienced this before implementing the SNV/SNP filter, where samples coming from the same population had a SNP signal so distinct from other samples, that it lead to the creation of a new signature during the initial NMF. This poses an even greater challenge in case of 5hmC, where the sites are never fully modified, and the unfiltered SNPs may greatly skew the average beta values, especially if calculating them from a range of blocks. One way to avoid is would be to use a database such as dbSNP (Kitts and Sherry 2011), collecting the most common SNPs across the genome allowing for their removal from our data. However, this approach also removes many sites which are not necessarily SNPs in our samples, greatly reducing the size of analysed input data. In this work, I presented a middle ground solution, where SNPs are calculated from each sample based on the count data and subsequently removed. This approach identified more fake methylation calls than dbSNP, substantially improving 5hmC data quality. I used a conservative approach also removing heterogenous SNPs, but in principle, one can adjust this method to appropriately handle heterogenous SNPs to preserve the unaffected allele.

In terms of identifying signatures and assigning appropriate weight coefficients to CpG blocks, there are multiple mathematical methods which have been used for the purpose of deconvolution of unknown samples. NMF has been widely used in modern biology in the context of gene expression, to describe tens of thousands of gene expression profiles in the terms of a small number of metagenes (Brunet et al. 2004), but also in one of the first reference-free methylation-based deconvolution studies (Houseman et al. 2016). Perhaps the most well-known use case is for the detection of hidden patterns of "mutational signatures" from mutation counts in trinucleotide contexts, which were assigned to known mutational processes, such as UV exposure or smoking (Nik-Zainal et al. 2012). In the cases of mutations

and gene expression, one deals with count data - any data point can be of a value from zero to thousands - the count of mutations or transcript is additive - the higher the number, the more information in the given data point. In the case of methylation, the data are bound tightly between 0 and 1. Using count data would not be appropriate in this context, because methylation is expressed as a proportion of modified to unmodified reads, so by definition it cannot be larger than 1. This presents several challenges in the context of NMF. Firstly, in biology, both the presence and the absence of something can be highly informative. This exactly happens with methylation, where having a beta value of zero is an information-bearing observation. In contrast, in NMF, a zero will be treated as a lack of signal/information, and it would not be selected in the model. As I can see on the heatmaps, the vast majority of the selected, high-variance blocks are hypomethylated rather than hypermethylated, which in the context of NMF would be presented as a 0 in the row full of 1s, and likely to be overlooked. This required us to "flip" the mostly methylated positions, so the now-informative site is a 1 in the row of 0s, which is then identifiable by NMF as an informative data point. NMF also favoured the values close to 0 and 1, assigning higher weights to extreme methylation states and omitting mid-methylation weights at regions such as enhancers.

The difference between NMF and traditional factorisation methods, such as PCA, is that NMF restricts the matrix representing the basis components and the matrix of mixture coefficients are constrained to have non-negative entries, allowing only for additive, and not subtractive, combinations. This is more difficult algorithmically, but allows for simple and intuitive interpretation of factors in NMF and allows for the basis components to overlap. For example, when NMF and PCA were applied to deconstruct images of faces, NMF produced images reminiscent of parts of faces such as eyes or noses. PCA, while being a simpler way to reduce dimensionality, produced a basis image that was composed of abstract constructs with negative and positive values that were difficult to interpret (Lee and Seung 1999). The advantage of NMF is the interpretability of the results. Despite the arbitrary unit of the "weights", I can clearly identify sites which tend to be the most important for a given signature and they do correspond to the points of reference from the input heatmaps. Additionally, in addition to manual selection of sites using variance filtering, NMF offers a flexible way to identify important blocks that can overlap with each other, reflecting shared developmental origins, or gain similarity of the tissues in question. Other publications and projects studying this matter focus on the very tissue-specific deconvolution application side of things and strongly

select tissue-specific sites, usually greatly limiting the number of CpGs left in the analysis. But this depends on the downstream application. As demonstrated in the validation, using the NMF-derived weights to perform least-squares regression classifies the samples accurately with respect to their tissue type.

In this work, 5hmC levels were used as a "second layer" of the analysis to understand the biological reasons for the selection of particular CpG blocks. I believe that 5hmC itself can contain additional information, and it would be most interesting to perform a similar deconvolution study based on 5hmC levels only. It is likely that the selection of tissue-specific sites would require some modifications of the method to account for the generally low 5mC patterns across the tissues. Additionally, as mentioned above, 5hmC maps are more fragile to variation in the form of SNPs, mutations, or sequencing errors, forcing a careful filtering process to remove any confounding effects.

In the third chapter, I showed the applicability of the tissue-specific methylation atlas in a cancer research setting. Fitting the 58,004 CpG blocks from the OAC trial samples to the NMF-derived coefficients revealed groups of tumours of distinct methylation profiles. The norm of residuals of the fit of each sample was strongly correlated with the tumour DNA content, providing a useful tool for examining tumour cellularity estimates. Furthermore, changes in the contribution of several signatures across time points were associated with different survival probabilities, suggesting that it could have a biomarker potential. The analysis was limited by several factors, including the quality of the TAPSβ atlas itself, especially in the context of high number of blocks with high weights shared across the signatures. Another limiting factor is the small number of available patient samples, especially ones sequenced at all three timepoints, significantly decreasing the power in any of the survival analyses. Despite these limitations, the deconvolution provided interesting insights into the context of the discussion of the origin of OAC as discussed before. The observed correlation between tumour purity and residual norms led me to hypothesize that tumour purity could be instrumental in discerning tumour-specific methylation patterns. I have attempted to isolate tumour signals using the purity and signature contribution values, but this was heavily limited by the nature of NNLS fit and by looking at only the blocks selected using the healthy samples in the reference. An interesting way to isolate tumour-specific methylation signals would be to perform reference-free deconvolution, as described in the other papers. This was beyond the scope of this thesis due to potential complications with genome segmentation caused by

the much greater heterogeneity of methylation signals. Differences in the levels of relative contributions of stomach and eosinophil signatures across treatment time points were found to be associated with altered survival probabilities, which is an interesting point to consider in the further analyses with more clinical samples and a more refined reference atlas.

In conclusion, this thesis described the development of a novel method to identify tissue-specific 5mC signals across 20 tissue types. The sites correspond to meaningful biological signals, associated with genes, open chromatin regions, enhancers, tissue-specific histone marks, and enrichment in the 5hmC signal. The method shows promising results in clinical applications, as demonstrated by the deconvolution of OAC samples and the identification of putative prognostic biomarkers.

6

Methods

Contents

6.1 Methylation datasets

6.1.1 TAPSβ atlas

Data processing

TAPSβ and CAPS sequencing on 72 samples was performed by Masato Inoue in Chunxiao Song's lab, and the raw sequencing data were processed by Robert Amess.

In short, genomic DNA from solid human tissues were obtained from OriGene and AMSBIO, while blood cells were isolated from whole blood of participating donors. Blood was separated into peripheral blood mononuclear cells (PBMCs) used to isolate monocytes, CD4+ and CD8+ T-cells, B-cells and NK cells using commerical kits. Granulocytes collected during the processing were used to isolate eosinophils, and the remaining cells were characterised as neutrophils. After isolation, cell purity was accessed with flow cytometry. TAPSβ was performed as previously described (Liu et al. 2021) with a few modifications on ligation and borane reduction. CAPS+ was performed as previously described (Xu et al. 2023). The sequenced libraries were processed with a standard pipeline. The reads were trimmed with Trim Galore! v0.6.5 and the quality was verified with FastQC v0.11.8, followed by an alignment with the hg38 genome with bwa-mem2 v2.2.1 (Krueger 2023; *FastQC* 2015; Vasimuddin et al. 2019). The technical replicates were merged and the duplicates marked with MarkDuplicates from Picard Tools v4.1.7.0 (*Picard Tools - By Broad Institute* 2023). Methylation calling was performed using MethylDackel v0.6.0. Only the methylation calls obtained in the CpG context and in autosomes were used in the remaining analysis (Ryan 2023).

The summary of all available samples is provided in Table A.1, which includes information on the samples removed due to poor coverage/processing. Three samples had to be excluded from the analysis due to quality issues and potential sample mislabeling.

- CD563504-Spleen – a mislabelled prostate sample
- CD563320-Spleen – poor quality control results and evidence of sample degradation
- CD563569-Liver – mislabelled liver cancer
- CD707005-Heart – poor coverage

- CD707007-Thymus – poor coverage
- CD603405-Bone Marrow – single sample not matching any other bone marrow methylation landscapes

For calculating the average methylation per CpG block, we used the following formula.

$$\beta = \frac{modC}{modC + unmodC}$$

List of all CpGs

A data frame containing the positions of the CpG sites within the hg38 genome was created using the BSgenome.Hsapiens.UCSC.hg38 package (Team TBD 2023).

Removal of blacklisted regions

CpGs overlapping regions listed in the ENCODE list of problematic regions were removed (Amemiya et al. 2019). This was done by reverse intersecting the full list of CpGs with the regions listed in the hg38-blacklist.v2.bed file available from the ENCODE repository using BEDTools (Quinlan and Hall 2010). Similarly, we removed hg38 centromere regions downloaded from the Genome Browser's Table Browser (Nassar et al. 2023).

SNV counting

To call possible SNVs in the TAPS atlas samples, I first created a data frame containing positions of all CpG sites in the hg38 genome using the BSgenome.Hsapiens.UCSC.hg38 package (Team TBD 2023). Then, I used a modified version of the alleleCount function on BAM files from each sample, using the CpG map as a template. The script reported alleles from the OB (original bottom - reads complementary to the original top strand) and OT (original top) strands separately (*alleleCount* 2023). Then, for each site, I calculated the proportion of alleles which differed from the expected allele under the assumption that the base opposite C in a CpG, modified or not, will always be a G. This means, that for the C in the given CpG, I considered alleles from the OB strand, and for G, I considered alleles from the OT strand (aligning to the original bottom strand, so to the C in CpG on the minus strand). If the proportion of Gs at each position was lower than 0.4, I assumed that the position is a SNV and not a cytosine modification. This

includes both homogenous and heterogenous SNVs, and is a rather conservative approach. Using this method, we created a bed file for each sample, indicating all sites which are potential SNVs. The positions were not filtered on the basis of coverage, because they are filtered elsewhere.

We used SNPdb v154 to test the overlap between the TAPS-derived SNVs and common SNPs in the population (Kitts and Sherry 2011).

Handling missing values

In cases where more than 25% of samples had missing data in a given block (i.e., no reads in the region covered by the block or insufficient coverage), we removed the entire block from the dataset. To maximise the number of blocks available for analysis, any remaining missing values were imputed using the median value of all other samples in the given block.

6.1.2 Loyfer atlas

The samples from the Loyfer et al. 2023 atlas were downloaded from GEO (accession no. GSE186458) as hg38 beta files. Beta is a binary format that was developed by the authors to be compatible with their wgbstools software, which has the option to convert .beta to .bed files. Because we had great difficulty installing the software in our computing environment, we obtained the index of CpG positions used to decipher the .beta files in the wgbstools software and converted the files manually. Each file contained the number of modified reads and total reads covering each CpG. We merged the file with the CpG index and calculated the number of unmodified reads by subtracting the number of modified reads from the total coverage. Additionally, we converted the read locations to match the 0-based coordinate system used in the remaining files throughout this thesis.

The positions of the top 1000 unmethylated blocks in the Loyfer et al. 2023 atlas were obtained from their supplementary table S4C. Block positions were converted from hg19 to hg38 using liftOver (Bioconductor 2021).

6.1.3 LUD2015-005 trial data

The research study procedures, sample collection, pre-sequencing sample handling and description of the assessment of clinical outcomes are described in Carroll et al. 2023. After isolation of genomic DNA, TAPS was performed as described

in Liu et al. 2019, by Paulina Siejka-Zielińska. Bioinformatics pipelines were performed by Robert Amess and Benjamin Schuster-Boeckler as described above in Methods 6.1.1.

6.2 Non-negative matrix factorisation

NMF algorithm

Non-negative matrix factorisation was performed in R using the NMF package (Gaujoux and Seoighe 2010). Prior to factorisation, the matrix containing average methylation values in each block was converted for all samples. If the median beta value of a block across all samples was larger than 0.3, all beta values were reversed so that:
if $median(blockBeta) > 0.3$ then $blockBeta < -1 - blockBeta$
The locations of each converted block were noted and stored along with the NMF result. Because the NMF algorithm does not accept zeros, we added a pseudo-count of 0.0001 to each cell. Brunet NMF algorithm was used in each experiment and the calculations were repeated 200 times each, with a set seed to ensure reproducibility. For the visualisations in the form of a contribution heatmap, the contributions of each signature to each tissue were first scaled to sum to 1.

High-weight block identification

To select the most important sites for each signature, we first obtained the weights of the blocks in each of the signatures used to reconstruct the NMF matrix. To avoid manual threshold setting, we applied the Expectation Maximisation (EM) algorithm using the normalmixEM R function from the mixtools package (Benaglia et al. 2010), ignoring the assumption of normality of both distributions. We set the cut-off point to be the arbitrary border between the two distributions, equal to $\mu - \frac{\sigma}{2}$. We treat all blocks above the cutoff as potentially informative.

NNLS regression

To fit new samples to previously obtained NMF coefficients, a new matrix was constructed reflecting the order of CpG blocks in the NMF, and the positions of blocks converted in the reference matrix. Each position converted in the reference was then converted in the same way ($blockBeta < -1 - blockBeta$). Then, each sample was fitted to the matrix of NMF coefficients using the lsqnonneg function from the pracma R package (Borchers 2021). The contributions of signatures and norm of residuals from each sample were directly extracted from the function output. Signature contributions were then scaled sample-wise to sum to one.

6.3 Annotations

6.3.1 Genes

For annotation of genes and their elements, the basic annotation GENCODE v43 was used (Frankish et al. 2019). All transcripts that are not included in the first version of the MANE Select set, which is an universal set of one transcript per protein-coding gene, were filtered (Morales et al. 2022).

6.3.2 CGIs

CGI annotations were downloaded using the package annotatr, using the "hg38_cpgs" set (Cavalcante and Sartor 2017).

6.3.3 Enhancers

Enhancer elements, their association with genes and tissue specificity were extracted from the GeneHancer database version 5.14 (Fishilevich et al. 2017). Genomic positions were adjusted to match 0-based standard of remaining files. The genes associated with each enhancer were filtered to match the genes included in the MANE database. Both elite and non-elite enhancers and associations were used in the analyses, unless stated otherwise. "Promoter" and "Promoter/Enhancer" types were excluded from the dataset to focus on distal enhancers specifically. Tissue-specific enhancer (ENCODE) and superenhancer (dbSUPER) sets were also obtained from GeneHancer.

6.3.4 Tissue-specific genes

Tissue/cell-type specific transcriptome was downloaded from The Human Protein Atlas project (Karlsson et al. 2021; *The Human Protein Atlas* 2023), and split into tissue-specific, cell-type-specific, blood cell-type-specific, and blood lineage-specific sets according to their description. The original specificity annotations were preserved. Genes which were not detected at all were removed, and genes of low tissue-specificity were annotated as "housekeeping" genes.

6.3.5 Other gene sets

Gene Ontology and KEGG Pathway gene sets were obtained from the Molecular Signatures Database (MSigDB) using the msigdbr package (Liberzon et al. 2011; Bhuva et al. 2021).

6.3.6 Overlapping genomic annotations

To compare the representation of genomic elements by CpG blocks, the regions were overlapped on a single CpG basis. A map with all CpG positions was intersected with bed files with the above annotations and with the information whether it belongs to blocks.

6.4 Enrichment analysis

All enrichment analyses were performed using the clusterProfiler R package using the enricher() function (Wu et al. 2021). For gene-based analyses the environment was set to represent only the genes overlapped by all blocks (instead of all genes). The same approach was used with enhancers. Benjamini-Hochberg correction was applied to adjust the p-values, and the p-value cutoff filter was set to 0.05.

6.5 ChIP-seq analysis

Histone modification ChIP-seq data were downloaded from the Roadmap Epigenomics database. H3K4me1 and H3K4me3 modifications were downloaded in narrowPeak format. H3K27me3, H3K36me3, and H3K9me3 states were downloaded as broadPeak files. The choice was guided by the localisation of each interaction, as suggested in (Landt et al. 2012; Sims et al. 2014). The files were lifted over from hg19 to hg38 using the liftOver tool (Hinrichs et al. 2006). We selected tissues corresponding to the tissues of our methylation atlas, the samples are summarised in table A.3.

The heatmap of peak over-representation in blocks important for distinct signatures was created by first overlapping the ChIP-seq peak locations from all reference tissues and blocks from all signatures. Then, contingency tables were created to count blocks present or not present in a ChIP-seq peak and to count whether it is included in a given signature or not. The overlaps were only counted if a minimum of 75% of the CpGs in each block overlapped with the peak. The chi-square test was performed on each contingency table and chi-square statistics were reported due to extremely low p values.

To calculate the odds ratios for the presence of peaks in blocks important for individual signatures, the ChIP-seq files were paired with their corresponding tissue signatures (as suggested by the over-representation analysis). Overlaps with chromatin marks were calculated as described above. The categories were defined as follows:

1. tissue-wide blocks - blocks which are described as important for more than 10 signatures
2. unique blocks - blocks important for a single signature
3. blood blocks - blocks where the ratio of them being important in blood to tissue signatures is >0.5 (thus enriching for blood-unique blocks)
4. tissue blocks - blocks where the ratio of them being important in tissue to blood signatures is >0.5 (thus enriching for tissue-unique blocks)

To calculate the over-representation of methylation states at each peak, for each signature we calculated the average beta value at each block from samples with the highest contribution of that signature (for example, from all B-cells for a B-cell-specific signature). Then, each block was calculated as methylated (beta value > 0.66), or unmethylated (the remaining sites). Fisher's exact test was performed on methylated/unmethylated/important/not important contingency tables.

The ChromHMM annotation of genomic regions from the 15-state model was downloaded from the Roadmap Epigenomics project as hg38-aligned bed files (Kundaje et al. 2015; Ernst and Kellis 2012). This model was trained on H3K4me3, H3K4me1, H3K36me3, H3K27me3 and H3K9me3 states of 60 epigenomes from the project. Only primary tissues and blood cell samples corresponding to cells in the atlas were used for the analyses, as summarised in the table A.3. All analyses were performed as described for the peak data.

6.6 Tumour DNA content estimation

Tumour DNA content was calculated from purity and ploidy under the assumption that the total DNA in a sample to be the sum of tumour DNA and normal DNA, such as:

$$tumourDNAcontent = \frac{tumourDNAamount}{totalDNAamount}$$

If there are n cells in the measured sample, then:

$$tumourDNAamount = n \times purity \times ploidy$$

$$normalDNAamount = n \times (1 - purity) \times 2$$

$$totalDNAamount = tumourDNAamount + normalDNAamount$$

Combining the equation gives us:

$$tumourDNAcontent = \frac{n \times purity \times ploidy}{n \times purity \times ploidy + n \times (1 - purity) \times 2}$$

Which can be simplified to:

$$tumourDNAcontent = \frac{purity \times ploidy}{purity \times ploidy + (1 - purity) \times 2}$$

The norm of residuals from each sample was obtained directly from the NNLS function output. A linear model was constructed for the association between tumour DNA content and norm of residuals using the following formula:

$$lm(tumourDNAcontent\ normOfResiduals)$$

The resulting linear model was used with the function approx() to calculate the expected tumour DNA content based on the norm of residuals.

The signature contributions were then scaled to reflect the potential tumour content, resulting in four types of possible contribution representation.

1. Unscaled contribution (direct output of the NMF and NNLS)
2. Contribution scaled to one (such that the sum of contributions from each sample sums to one, this is the version mostly used/visualised throughout the thesis)
3. Unscaled contribution corrected for tumour DNA content ($contribution * SumOfContributions * (1 - \%tumourDNAcontent)$). Tumour DNA content in this case is calculated as $SumOfContributions * \%tumourDNAcontent$.
4. Scaled contribution corrected for tumour DNA content ($scaledContribution * (1 - \%tumourDNAcontent)$).

6.7 Survival analysis

Survival analyses were performed and presented using the survival and survminer R packages as in the following code:

$$survfit(Surv(overallSurvival, status)\ contributionClass)$$

The $contributionClass$ logical variables represented the signature contribution being above/below the median in terms of basal signature contribution, which was calculated across all samples within each contribution category. Differences in contributions were calculated by subtracting the previous timepoint, such that a positive result indicated an increase of signature contribution across the timepoints, and negative result indicated a decrease. Samples with differences of 0 were excluded from the survival analysis.

6.8 Computational environments

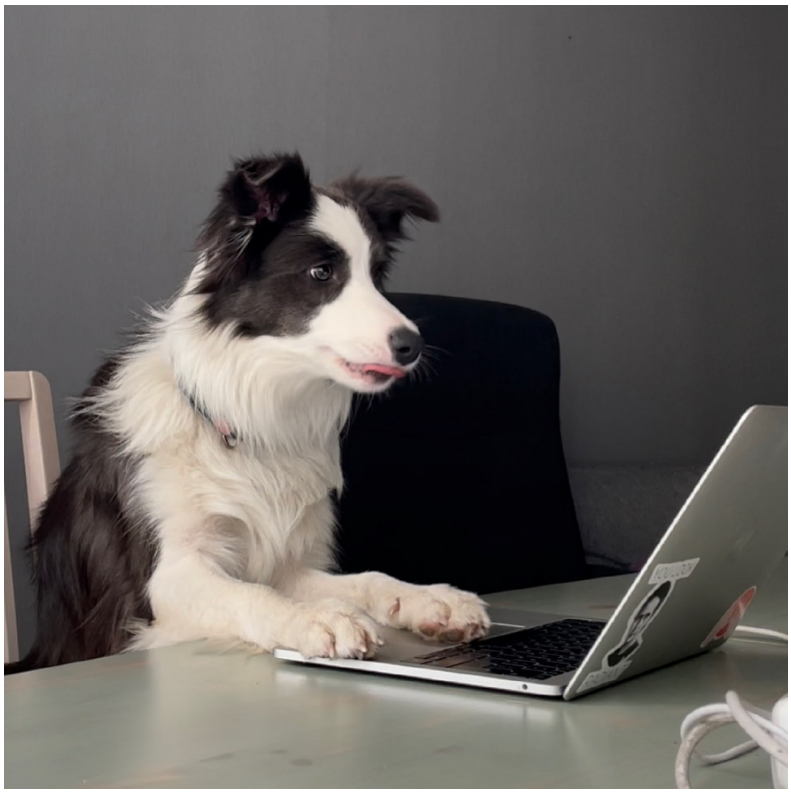
All analyses were conducted on the Biomedical Research Computing (BMRC) high-performance cluster at the Big Data Institute.

All analyses, unless otherwise stated, were performed in R version 4.1.0. Analysis-specific packages are described in relevant sessions. Packages used in multiple sections include: ComplexHeatmap (Gu et al. 2016), tidyverse (Wickham et al. 2019), data.table (Dowle and Srinivasan 2023), GenomicRanges (Lawrence et al. 2013).

7

Special acknowledgements

Special acknowledgement to Luna, who almost wrote some code for the project.



Appendices

A

Datasets used in the thesis

Np.	Sample ID	Tissue/cell type	Donor ID
1	UKVAC-140_B-cells	B-cells	UKVAC-140
2	UKVAC-003-6_B-cells	B-cells	UKVAC-003-6
3	UKVAC-001-5_B-cells	B-cells	UKVAC-001-5
4	UKVAC-049-3_B-cells	B-cells	UKVAC-049-3
5	C603405_Bone-Marrow*	Bone-Marrow	C603405
6	C505016_Brain	Brain	C505016
7	C707004_Brain	Brain	C707004
8	CD563137_Breast	Breast	CD563137
9	CD564068_Breast	Breast	CD564068
10	CD563304_Breast	Breast	CD563304
11	CD564368_Breast	Breast	CD564368
12	UKVAC-140_CD4-T-cells	CD4-T-cells	UKVAC-140
13	UKVAC-049-3_CD4-T-cells	CD4-T-cells	UKVAC-049-3
14	UKVAC-003-6_CD4-T-cells	CD4-T-cells	UKVAC-003-6
15	UKVAC-001-5_CD4-T-cells	CD4-T-cells	UKVAC-001-5
16	UKVAC-049-3_CD8-T-cells	CD8-T-cells	UKVAC-049-3
17	UKVAC-140_CD8-T-cells	CD8-T-cells	UKVAC-140
18	UKVAC-001-5_CD8-T-cells	CD8-T-cells	UKVAC-001-5
19	UKVAC-003-6_CD8-T-cells	CD8-T-cells	UKVAC-003-6
20	CD564159_Colon	Colon	CD564159
21	CD563663_Colon	Colon	CD563663
22	CD563419_Colon	Colon	CD563419
23	CD565189_Colon	Colon	CD565189
24	UKVAC-049-3_Eosinophils	Eosinophils	UKVAC-049-3
25	UKVAC-001-5_Eosinophils	Eosinophils	UKVAC-001-5
26	UKVAC-003-6_Eosinophils	Eosinophils	UKVAC-003-6
27	UKVAC-140_Eosinophils	Eosinophils	UKVAC-140
28	CD563678_Esophagus	Esophagus	CD563678
29	CD565252_Esophagus	Esophagus	CD565252
30	CD564986_Esophagus	Esophagus	CD564986
31	CD565136_Esophagus	Esophagus	CD565136
32	C707005_Heart*	Heart	C707005
33	C707006_Heart	Heart	C707006
34	CD564241_Kidney	Kidney	CD564241
35	CD563984_Kidney	Kidney	CD563984
36	CD563778_Kidney	Kidney	CD563778
37	CD563776_Kidney	Kidney	CD563776
38	CD563569_Liver*	Liver	CD563569
39	CD564797_Liver	Liver	CD564797
40	CD564082_Liver	Liver	CD564082
41	CD563257_Liver	Liver	CD563257
42	CD564901_Lung	Lung	CD564901
43	CD564511_Lung	Lung	CD564511
44	CD563930_Lung	Lung	CD563930

45	CD563854_Lung	Lung	CD563854
46	UKVAC-003-6_Monocytes	Monocytes	UKVAC-003-6
47	UKVAC-001-5_Monocytes	Monocytes	UKVAC-001-5
48	UKVAC-140_Monocytes	Monocytes	UKVAC-140
49	UKVAC-049-3_Monocytes	Monocytes	UKVAC-049-3
50	UKVAC-049-3_Neutrophils	Neutrophils	UKVAC-049-3
51	UKVAC-001-5_Neutrophils	Neutrophils	UKVAC-001-5
52	UKVAC-003-6_Neutrophils	Neutrophils	UKVAC-003-6
53	UKVAC-140_Neutrophils	Neutrophils	UKVAC-140
54	UKVAC-003-6_NK-cells	NK-cells	UKVAC-003-6
55	UKVAC-001-5_NK-cells	NK-cells	UKVAC-001-5
56	UKVAC-140_NK-cells	NK-cells	UKVAC-140
57	UKVAC-049-3_NK-cells	NK-cells	UKVAC-049-3
58	CD564191_Ovary	Ovary	CD564191
59	CD563544_Ovary	Ovary	CD563544
60	CD564295_Ovary	Ovary	CD564295
61	CD564404_Pancreas	Pancreas	CD564404
62	CD565341_Pancreas	Pancreas	CD565341
63	CD564844_Pancreas	Pancreas	CD564844
64	CD564011_Pancreas	Pancreas	CD564011
65	CD564242_Prostate	Prostate	CD564242
66	CD563610_Prostate	Prostate	CD563610
67	CD563685_Prostate	Prostate	CD563685
68	CD563267_Prostate	Prostate	CD563267
69	CD565017_Spleen	Spleen	CD565017
70	CD563880_Spleen	Spleen	CD563880
71	CD563504_Spleen*	Spleen	CD563504
72	CD563320_Spleen*	Spleen	CD563320
73	CD563430_Stomach	Stomach	CD563430
74	CD564596_Stomach	Stomach	CD564596
75	CD563162_Stomach	Stomach	CD563162
76	CD565042_Stomach	Stomach	CD565042
77	C707007_Thymus*	Thymus	C707007

Table A.1: TAPS atlas datasets
Datasets used in the thesis. Samples with asterisks (*) were removed due to technical concerns. The "Tissue/cell type" column corresponds to the groups within which samples were merged together at the block selection step of the analysis.

Np.	**sample_title**	**tissue**
1	Adipocytes-Z000000T7	abdominal_subcut_adipocytes
2	Adipocytes-Z000000T9	abdominal_subcut_adipocytes
3	Adipocytes-Z000000T5	abdominal_subcut_adipocytes
4	Aorta-Endothelium-Z00000422	aorta_endothelium
5	Aorta-Endothelium-Z0000043G	aorta_endothelium
6	Saphenous-Vein-Endothelium-Z000000RM	vascular_saphenous_endothelium
7	Saphenous-Vein-Endothelium-Z000000S7	vascular_saphenous_endothelium
8	Saphenous-Vein-Endothelium-Z000000SB	vascular_saphenous_endothelium
9	Kidney-Glomerular-Endothelium-Z000000Q5	kidney_glomerular_endothelium
10	Kidney-Glomerular-Endothelium-Z00000443	kidney_glomerular_endothelium
11	Kidney-Glomerular-Endothelium-Z0000045J	kidney_glomerular_endothelium
12	Kidney-Tubular-Endothelium-Z000000PX	kidney_tubular_endothelium
13	Kidney-Tubular-Endothelium-Z000000Q3	kidney_tubular_endothelium
14	Kidney-Tubular-Endothelium-Z0000042R	kidney_tubular_endothelium
15	Lung-Alveolar-Endothelium-Z000000Q1	lung_alveolar_endothelium
16	Lung-Alveolar-Endothelium-Z000000QK	lung_alveolar_endothelium
17	Lung-Alveolar-Endothelium-Z0000045H	lung_alveolar_endothelium
18	Pancreas-Endothelium-Z0000042D	pancreas_endothelium
19	Pancreas-Endothelium-Z0000042X	pancreas_endothelium
20	Pancreas-Endothelium-Z00000430	pancreas_endothelium
21	Pancreas-Islet-Endothelium-Z0000042Y	pancreas_endothelium
22	Colon-Fibroblasts-Z0000042A	colon_fibroblast
23	Colon-Fibroblasts-Z0000042C	colon_fibroblast
24	Heart-Fibroblasts-Z0000043R	heart_fibroblast
25	Heart-Fibroblasts-Z0000041V	heart_fibroblast
26	Heart-Fibroblasts-Z0000041W	heart_fibroblast
27	Heart-Fibroblasts-Z0000041X	heart_fibroblast
28	Skeletal-Muscle-Z00000427	skeletal_muscle_striated_muscle
29	Skeletal-Muscle-Z00000429	skeletal_muscle_striated_muscle
30	Heart-Cardiomyocyte-Z0000044G	heart_cardiomyocyte
31	Heart-Cardiomyocyte-Z0000044K	heart_cardiomyocyte
32	Heart-Cardiomyocyte-Z0000044N	heart_cardiomyocyte
33	Heart-Cardiomyocyte-Z0000044P	heart_cardiomyocyte
34	Heart-Cardiomyocyte-Z0000044Q	heart_cardiomyocyte
35	Heart-Cardiomyocyte-Z0000044R	heart_cardiomyocyte
36	Oligodendrocytes-Z000000TK	brain_oligodendrocytes
37	Oligodendrocytes-Z0000042E	brain_oligodendrocytes
38	Oligodendrocytes-Z0000042L	brain_oligodendrocytes
39	Oligodendrocytes-Z0000042N	brain_oligodendrocytes
40	Cortex-Neuron-Z000000TF	brain_neuronal
41	Neuron-Z000000TH	brain_neuronal
42	Cortex-Neuron-Z0000042F	brain_neuronal
43	Cortex-Neuron-Z0000042H	brain_neuronal
44	Cortex-Neuron-Z0000042J	brain_neuronal

45	Cortex-Neuron-Z0000042M	brain_neuronal
46	Cortex-Neuron-Z0000042P	brain_neuronal
47	Cortex-Neuron-Z0000042K	brain_neuronal
48	Cerebellum-Neuron-Z000000TB	brain_neuronal
49	Cortex-Neuron-Z000000TD	brain_neuronal
50	Liver-Hepatocytes-Z000000R3	liver_hepatocyte
51	Liver-Hepatocytes-Z000000T3	liver_hepatocyte
52	Liver-Hepatocytes-Z0000043Q	liver_hepatocyte
53	Liver-Hepatocytes-Z0000044H	liver_hepatocyte
54	Liver-Hepatocytes-Z0000044M	liver_hepatocyte
55	Liver-Hepatocytes-Z00000431	liver_hepatocyte
56	Pancreas-Duct-Z0000043T	pancreas_duct
57	Pancreas-Duct-Z0000043U	pancreas_duct
58	Pancreas-Duct-Z0000043V	pancreas_duct
59	Pancreas-Duct-Z000000QZ	pancreas_duct
60	Pancreas-Acinar-Z000000QX	pancreas_acinar
61	Pancreas-Acinar-Z0000043W	pancreas_acinar
62	Pancreas-Acinar-Z0000043X	pancreas_acinar
63	Pancreas-Acinar-Z0000043Y	pancreas_acinar
64	Pancreas-Delta-Z00000451	pancreas_delta
65	Pancreas-Delta-Z00000454	pancreas_delta
66	Pancreas-Delta-Z00000457	pancreas_delta
67	Pancreas-Beta-Z00000452	pancreas_beta
68	Pancreas-Beta-Z00000455	pancreas_beta
69	Pancreas-Beta-Z00000458	pancreas_beta
70	Pancreas-Alpha-Z00000453	pancreas_alpha
71	Pancreas-Alpha-Z00000456	pancreas_alpha
72	Pancreas-Alpha-Z00000459	pancreas_alpha
73	Kidney-Glomerular-Epithelium-Z0000045K	kidney_glomerular_epithelium
74	Kidney-Glomerular-Epithelium-Z0000045L	kidney_glomerular_epithelium
75	Kidney-Tubular-Epithelium-Z000000QH	kidney_tubular_epithelium
76	Kidney-Tubular-Epithelium-Z0000043Z	kidney_tubular_epithelium
77	Kidney-Tubular-Epithelium-Z00000440	kidney_tubular_epithelium
78	Kidney-Glomerular-Podocytes-Z0000042W	kidney_glomerular_podocyte
79	Kidney-Glomerular-Podocytes-Z00000441	kidney_glomerular_podocyte
80	Kidney-Glomerular-Podocytes-Z00000442	kidney_glomerular_podocyte
81	Thyroid-Epithelium-Z0000042S	thyroid_epithelium
82	Thyroid-Epithelium-Z0000042T	thyroid_epithelium
83	Thyroid-Epithelium-Z0000042U	thyroid_epithelium
84	Fallopian-Epithelium-Z000000Q7	fallopien_tubes_epithelium
85	Fallopian-Epithelium-Z000000S9	fallopien_tubes_epithelium
86	Fallopian-Epithelium-Z000000UV	fallopien_tubes_epithelium
87	Endometrium-Epithelium-Z00000434	endometrium_epithelium
88	Endometrium-Epithelium-Z00000435	endometrium_epithelium
89	Endometrium-Epithelium-Z0000043S	endometrium_epithelium

90	Bone_marrow-Erythrocyte_progenitors-Z000000RF	bone_marrow_erythrocyte_progeni
91	Bone_marrow-Erythrocyte_progenitors-Z000000RH	bone_marrow_erythrocyte_progeni
92	Bone_marrow-Erythrocyte_progenitors-Z000000RK	bone_marrow_erythrocyte_progeni
93	Blood-T-CD3-Z000000TV	blood_t_cd3_cells
94	Blood-T-CD3-Z000000UP	blood_t_cd3_cells
95	Blood-T-CD4-Z000000TT	blood_t_helpercd4_cells
96	Blood-T-CD4-Z000000U7	blood_t_helpercd4_cells
97	Blood-T-CD4-Z000000UM	blood_t_helpercd4_cells
98	Blood-T-CD8-Z000000TR	blood_t_cytotoxic_cd8_cells
99	Blood-T-CD8-Z000000U5	blood_t_cytotoxic_cd8_cells
100	Blood-T-CD8-Z000000UK	blood_t_cytotoxic_cd8_cells
101	Blood-T-CenMem-CD4-Z00000417	blood_t_central_memory_cd4
102	Blood-T-CenMem-CD4-Z0000041D	blood_t_central_memory_cd4
103	Blood-T-CenMem-CD4-Z0000041N	blood_t_central_memory_cd4
104	Blood-T-Eff-CD8-Z00000419	blood_t_effector_cell_cd8
105	Blood-T-Eff-CD8-Z0000041F	blood_t_effector_cell_cd8
106	Blood-T-Eff-CD8-Z0000041Q	blood_t_effector_cell_cd8
107	Blood-T-EffMem-CD4-Z00000416	blood_t_effector_memory_cd4
108	Blood-T-EffMem-CD4-Z0000041C	blood_t_effector_memory_cd4
109	Blood-T-EffMem-CD4-Z0000041M	blood_t_effector_memory_cd4
110	Blood-T-EffMem-CD8-Z0000041A	blood_t_effector_memory_cd8
111	Blood-T-EffMem-CD8-Z0000041G	blood_t_effector_memory_cd8
112	Blood-T-Naive-CD8-Z0000041B	blood_naive_t_cells_cd8
113	Blood-T-Naive-CD8-Z0000041H	blood_naive_t_cells_cd8
114	Blood-NK-Z000000TM	blood_nk
115	Blood-NK-Z000000U1	blood_nk
116	Blood-NK-Z000000UF	blood_nk
117	Blood-Monocytes-Z000000TP	blood_monocytes
118	Blood-Monocytes-Z000000U3	blood_monocytes
119	Blood-Monocytes-Z000000UH	blood_monocytes
120	Colon-Macrophages-Z00000444	colon_macrophages
121	Colon-Macrophages-Z00000446	colon_macrophages
122	Lung-Alveolar-Macrophages-Z00000448	lung_alveolar_macrophages
123	Lung-Alveolar-Macrophages-Z0000044C	lung_alveolar_macrophages
124	Lung-Interstitial-Macrophages-Z00000447	lung_interstitial_macrophages
125	Lung-Interstitial-Macrophages-Z0000044D	lung_interstitial_macrophages
126	Lung-Interstitial-Macrophages-Z0000044E	lung_interstitial_macrophages
127	Blood-Granulocytes-Z000000TZ	blood_granulocytes
128	Blood-Granulocytes-Z000000UD	blood_granulocytes
129	Blood-Granulocytes-Z000000UT	blood_granulocytes
130	Blood-B-Z000000TX	blood_b_cells
131	Blood-B-Z000000UB	blood_b_cells
132	Blood-B-Z000000UR	blood_b_cells
133	Blood-B-Mem-Z0000041J	blood_memory_b_cells
134	Blood-B-Mem-Z0000041K	blood_memory_b_cells

135	Tonsil-Palatine-Epithelium-Z000000QF	tonsil_palatine_epithelium
136	Tonsil-Palatine-Epithelium-Z000000RP	tonsil_palatine_epithelium
137	Tonsil-Palatine-Epithelium-Z000000RR	tonsil_palatine_epithelium
138	Tonsil-Pharyngeal-Epithelium-Z000000Q9	tonsil_pharyngeal_epithelium
139	Tonsil-Pharyngeal-Epithelium-Z000000S1	tonsil_pharyngeal_epithelium
140	Tongue-Epithelium-Z000000QV	tongue_epithelium
141	Tongue-Epithelium-Z00000449	tongue_epithelium
142	Tongue-Epithelium-Z0000044F	tongue_epithelium
143	Tongue_base-Epithelium-Z0000044B	tongue_epithelium
144	Esophagus-Epithelium-Z000000PZ	esophagus_epithelium
145	Esophagus-Epithelium-Z00000426	esophagus_epithelium
146	Lung-Bronchus-Epithelium-Z000000QD	lung_bronchus_epithelium
147	Lung-Bronchus-Epithelium-Z000000RZ	lung_bronchus_epithelium
148	Lung-Bronchus-Epithelium-Z000000S5	lung_bronchus_epithelium
149	Prostate-Epithelium-Z000000RV	prostate_epithelium
150	Prostate-Epithelium-Z000000S3	prostate_epithelium
151	Prostate-Epithelium-Z0000045F	prostate_epithelium
152	Prostate-Epithelium-Z0000045G	prostate_epithelium
153	Bladder-Epithelium-Z000000QM	bladder_epithelium
154	Bladder-Epithelium-Z000000QP	bladder_epithelium
155	Bladder-Epithelium-Z0000043F	bladder_epithelium
156	Bladder-Epithelium-Z0000044L	bladder_epithelium
157	Bladder-Epithelium-Z00000450	bladder_epithelium
158	Breast-Luminal-Epithelium-Z000000V2	breast_luminal_epithelial
159	Breast-Luminal-Epithelium-Z000000VJ	breast_luminal_epithelial
160	Breast-Luminal-Epithelium-Z000000VN	breast_luminal_epithelial
161	Breast-Basal-Epithelium-Z000000V6	breast_basal_epithelial
162	Breast-Basal-Epithelium-Z000000VG	breast_basal_epithelial
163	Breast-Basal-Epithelium-Z000000VL	breast_basal_epithelial
164	Breast-Basal-Epithelium-Z0000043E	breast_basal_epithelial
165	Lung-Alveolar-Epithelium-Z000000T1	lung_alveolar_epithelium
166	Lung-Alveolar-Epithelium-Z000000VC	lung_alveolar_epithelium
167	Lung-Alveolar-Epithelium-Z000000VE	lung_alveolar_epithelium
168	Gastric-fundus-Epithelium-Z000000RX	gastric_fundus_epithelium
169	Gastric-fundus-Epithelium-Z000000SK	gastric_fundus_epithelium
170	Gastric-fundus-Epithelium-Z000000SV	gastric_fundus_epithelium
171	Gastric-body-Epithelium-Z000000SD	gastric_body_epithelium
172	Gastric-body-Epithelium-Z000000SM	gastric_body_epithelium
173	Gastric-body-Epithelium-Z000000ST	gastric_body_epithelium
174	Gastric-antrum-Epithelium-Z000000SF	gastric_antrum_epithelium
175	Gastric-antrum-Epithelium-Z000000SP	gastric_antrum_epithelium
176	Gastric-antrum-Epithelium-Z000000SR	gastric_antrum_epithelium
177	Gastric-antrum-Endocrine-Z00000437	gastric_endocrine
178	Gastric-antrum-Endocrine-Z00000438	gastric_endocrine
179	Colon-Right-Epithelium-Z000000V0	colon_epithelium

180	Colon-Right-Epithelium-Z000000V8	colon_epithelium
181	Colon-Right-Endocrine-Z0000044S	colon_endocrine
182	Colon-Left-Epithelium-Z000000VA	colon_epithelium
183	Colon-Left-Endocrine-Z0000044J	colon_endocrine
184	Colon-Left-Endocrine-Z0000044T	colon_endocrine
185	Colon-Left-Epithelium-Z0000043B	colon_epithelium
186	Colon-Left-Epithelium-Z0000043C	colon_epithelium
187	Small-int-Epithelium-Z000000RT	small_intestine_epithelium
188	Small-int-Epithelium-Z000000UW	small_intestine_epithelium
189	Small-int-Epithelium-Z000000UY	small_intestine_epithelium

Table A.2: Loyfer atlas datasets
Datasets from Loyfer et al. 2023 used in the thesis, together with their tissue assignments.

epigenome ID	epigenome name	simplified name
E038	CD4 Naive Primary Cells	CD4 T-cells
E047	CD8 Naive Primary Cells	CD8 T-cells
E029	CD14 Primary Cells	Monocytes
E032	CD19 Primary Cells	B cells
E046	CD56 Primary Cells	NK cells
E030	CD15 Primary Cells	Neutrophils
E027	Breast Myoepithelial Cells	Breast Myoepithelial
E112	Thymus	Thymus
E104	Rigtht Atrium	Right Atrium
E109	Small Intestine	Small Intestine
E106	Sigmoid Colon	Sigmoid Colon
E079	Oesophagus	Oesophagus
E083	Fetal Heart	Fetal Heart
E086	Fetal Kidney	Kidney Kidney
E088	Fetal Lung	Fetal Lung
E095	Left Ventricle	Left Ventricle
E097	Ovary	Ovary
E066	Adult Liver	Liver
E098	Pancreas	Pancreas
E096	Lung	Lung
E102	Rectal Mucosa Donor 31	Rectal Mucosa
E110	Stomach Mucosa	Stomach
E113	Spleen	Spleen

Table A.3: ChIP-seq datasets
Descripion.

B

Supplementary figures

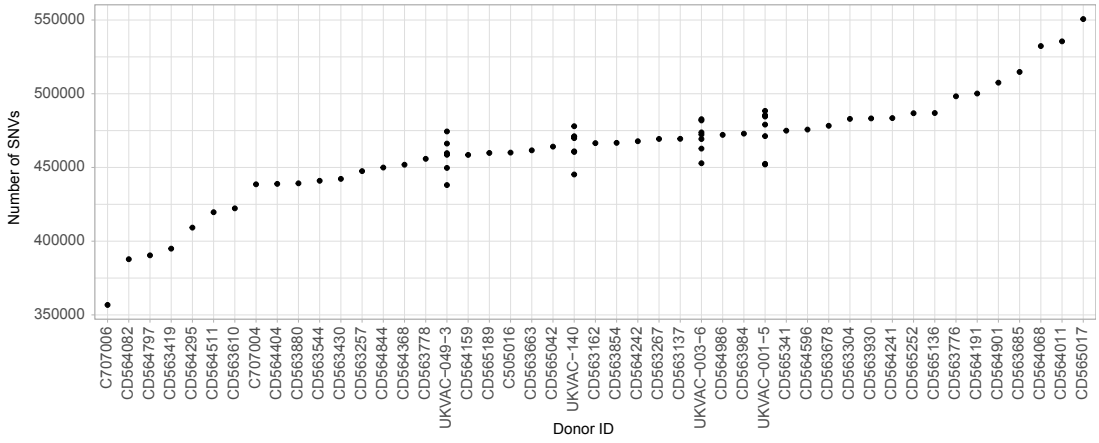


Figure B.1: SNV per donor in the TAPSβ atlas
Number of SNVs identified using the allele counting method described in 6.1.1. Results are not adjusted for coverage.

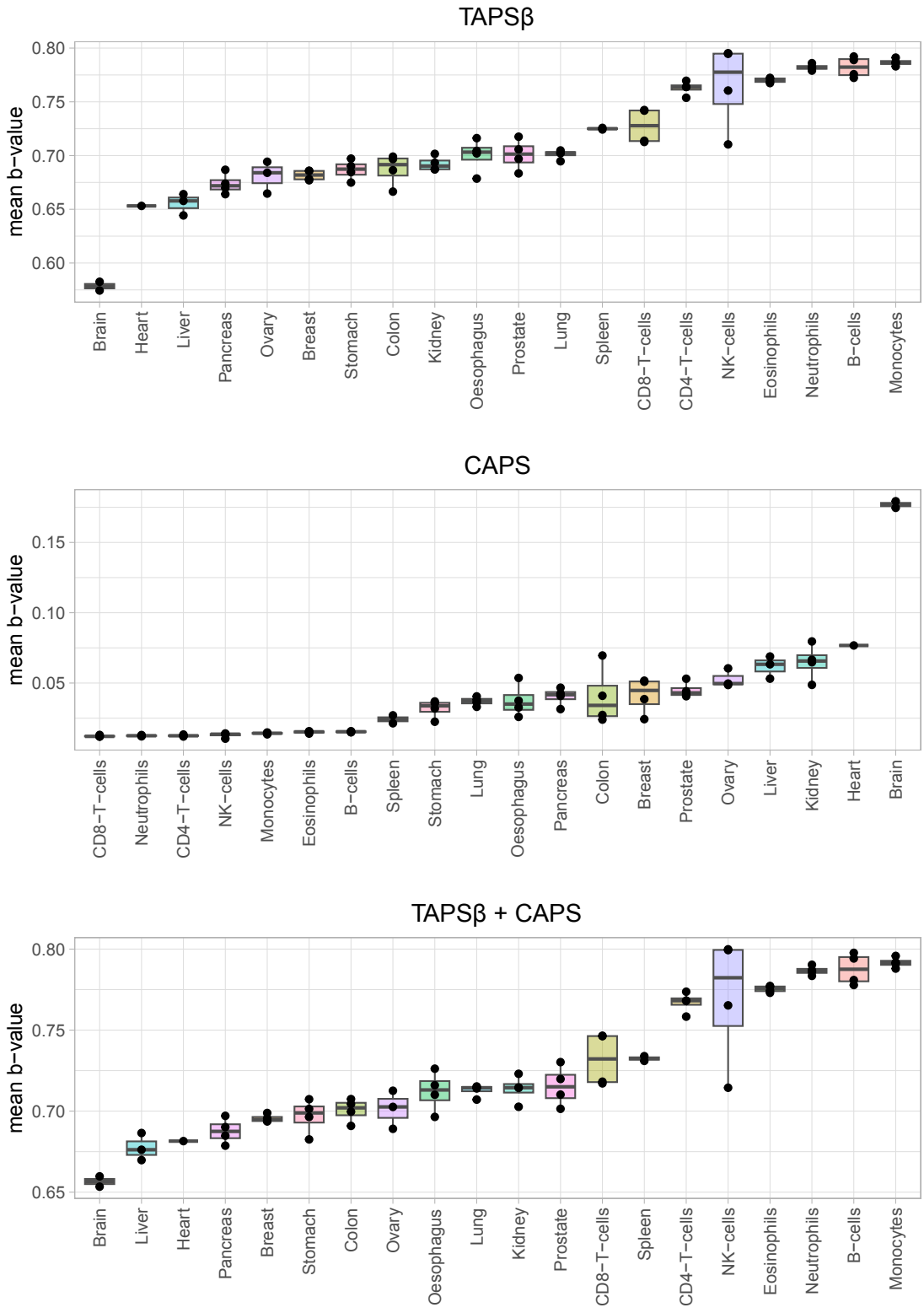


Figure B.2: Average modification values in TAPSβ and CAPS atlases
The average modification values were calculated from all samples above minimal coverage cutoff. The last panel illustrates TAPSβ atlas values, where some the unmodified reads are turned into modified, depending on the corresponding 5hmC beta value from CAPS.

References

Abe, Y., T. Koike, K. Iijima, A. Imatani, K. Ishida, T. Yuki, G. Miyata and T. Shimosegawa (2011). 'Esophageal Adenocarcinoma Developing after Eradication of Helicobacter Pylori'. *Case Reports in Gastroenterology* 5.2, pp. 355–360. DOI: 10.1159/000329878.

Alexandrov, L. B. et al. (2013). 'Signatures of Mutational Processes in Human Cancer'. *Nature* 500.7463 (7463), pp. 415–421. DOI: 10.1038/nature12477.

alleleCount (2023). URL: http://cancerit.github.io/alleleCount/#allelecount (visited on 02/08/2023).

Amemiya, H. M., A. Kundaje and A. P. Boyle (2019). 'The ENCODE Blacklist: Identification of Problematic Regions of the Genome'. *Scientific Reports* 9.1 (1), p. 9354. DOI: 10.1038/s41598-019-45839-z.

Anastasiadi, D., A. Esteve-Codina and F. Piferrer (2018). 'Consistent Inverse Correlation between DNA Methylation of the First Intron and Gene Expression across Tissues and Species'. *Epigenetics & Chromatin* 11.1, p. 37. DOI: 10.1186/s13072-018-0205-1.

Bachman, M., S. Uribe-Lewis, X. Yang, M. Williams, A. Murrell and S. Balasubramanian (2014). '5-Hydroxymethylcytosine Is a Predominantly Stable DNA Modification'. *Nature Chemistry* 6.12 (12), pp. 1049–1055. DOI: 10.1038/nchem.2064.

Ball, M. P., J. B. Li, Y. Gao, J.-H. Lee, E. M. LeProust, I.-H. Park, B. Xie, G. Q. Daley and G. M. Church (2009). 'Targeted and Genome-Scale Strategies Reveal Gene-Body Methylation Signatures in Human Cells'. *Nature Biotechnology* 27.4, pp. 361–368. DOI: 10.1038/nbt.1533.

Banerji, J., S. Rusconi and W. Schaffner (1981). 'Expression of a Beta-Globin Gene Is Enhanced by Remote SV40 DNA Sequences'. *Cell* 27 (2 Pt 1), pp. 299–308. DOI: 10.1016/0092-8674(81)90413-x.

Baubec, T., D. F. Colombo, C. Wirbelauer, J. Schmidt, L. Burger, A. R. Krebs, A. Akalin and D. Schübeler (2015). 'Genomic Profiling of DNA Methyltransferases Reveals a Role for DNMT3B in Genic Methylation'. *Nature* 520.7546 (7546), pp. 243–247. DOI: 10.1038/nature14176.

Beck, S. and V. K. Rakyan (2008). 'The Methylome: Approaches for Global DNA Methylation Profiling'. *Trends in Genetics* 24.5, pp. 231–237. DOI: 10.1016/j.tig.2008.01.006.

Bellacosa, A. and A. C. Drohat (2015). 'Role of Base Excision Repair in Maintaining the Genetic and Epigenetic Integrity of CpG Sites'. *DNA repair* 32, pp. 33–42. DOI: 10.1016/j.dnarep.2015.04.011.

Ben-Hattar, J. and J. Jiricny (1988). 'Methylation of Single CpG Dinucleotides within a Promoter Element of the Herpes Simplex Virus Tk Gene Reduces Its Transcription in Vivo'. *Gene* 65.2, pp. 219–227. DOI: 10.1016/0378-1119(88)90458-1.

Benaglia, T., D. Chauveau, D. R. Hunter and D. S. Young (2010). 'Mixtools: An R Package for Analyzing Mixture Models'. *Journal of Statistical Software* 32, pp. 1–29. DOI: 10.18637/jss.v032.i06.

Bernstein, B. E., A. Meissner and E. S. Lander (2007). 'The Mammalian Epigenome'. *Cell* 128.4, pp. 669–681. DOI: 10.1016/j.cell.2007.01.033.

Bestor, T. H. (2000). 'The DNA Methyltransferases of Mammals'. *Human Molecular Genetics* 9.16, pp. 2395–2402. DOI: 10.1093/hmg/9.16.2395.

Bhattacharyya, S. et al. (2017). 'Altered Hydroxymethylation Is Seen at Regulatory Regions in Pancreatic Cancer and Regulates Oncogenic Pathways'. *Genome Research* 27.11, pp. 1830–1842. DOI: 10.1101/gr.222794.117.

Bhuva, D. D., G. K. Smyth and A. Garnham (2021). *Msigdb: An ExperimentHub Package for the Molecular Signatures Database (MSigDB)*. manual.

Bibikova, M. et al. (2011). 'High Density DNA Methylation Array with Single CpG Site Resolution'. *Genomics*. New Genomic Technologies and Applications 98.4, pp. 288–295. DOI: 10.1016/j.ygeno.2011.07.007.

Bioconductor (2021). *liftOver: Changing Genomic Coordinate Systems with Rtracklayer::liftOver*. manual.

Bird, A. P. (1980). 'DNA Methylation and the Frequency of CpG in Animal DNA.' *Nucleic Acids Research* 8.7, pp. 1499–1504.

– (1984). 'DNA Methylation versus Gene Expression'. *Journal of Embryology and Experimental Morphology* 83 Suppl, pp. 31–40.

Bird, A. (2002). 'DNA Methylation Patterns and Epigenetic Memory'. *Genes & Development* 16.1, pp. 6–21. DOI: 10.1101/gad.947102.

Bird, A. P. and M. H. Taggart (1980). 'Variable Patterns of Total DNA and rDNA Methylation in Animals'. *Nucleic Acids Research* 8.7, pp. 1485–1497. DOI: 10.1093/nar/8.7.1485.

Blokzijl, F. et al. (2016). 'Tissue-Specific Mutation Accumulation in Human Adult Stem Cells during Life'. *Nature* 538.7624 (7624), pp. 260–264. DOI: 10.1038/nature19768.

Bodega, B. and V. Orlando (2014). 'Repetitive Elements Dynamics in Cell Identity Programming, Maintenance and Disease'. *Current Opinion in Cell Biology* 31, pp. 67–73. DOI: 10.1016/j.ceb.2014.09.002.

Borchers, H. W. (2021). *Pracma: Practical Numerical Math Functions*. manual.

Bostick, M., J. K. Kim, P.-O. Estève, A. Clark, S. Pradhan and S. E. Jacobsen (2007). 'UHRF1 Plays a Role in Maintaining DNA Methylation in Mammalian Cells'. *Science* 317.5845, pp. 1760–1764. DOI: 10.1126/science.1147939.

Bray, F., J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre and A. Jemal (2018). 'Global Cancer Statistics 2018: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries'. *CA: a cancer journal for clinicians* 68.6, pp. 394–424. DOI: 10.3322/caac.21492.

Brenet, F., M. Moh, P. Funk, E. Feierstein, A. J. Viale, N. D. Socci and J. M. Scandura (2011). 'DNA Methylation of the First Exon Is Tightly Linked to Transcriptional Silencing'. *PloS One* 6.1, e14524. DOI: 10.1371/journal.pone.0014524.

Brunet, J.-P., P. Tamayo, T. R. Golub and J. P. Mesirov (2004). 'Metagenes and Molecular Pattern Discovery Using Matrix Factorization'. *Proceedings of the National Academy of Sciences* 101.12, pp. 4164–4169. DOI: 10.1073/pnas.0308531101.

Cancer Genome Atlas Research Network (2012). 'Comprehensive Genomic Characterization of Squamous Cell Lung Cancers'. *Nature* 489.7417, pp. 519–525. DOI: 10.1038/nature11404.

Capuano, F., M. Mülleder, R. Kok, H. J. Blom and M. Ralser (2014). 'Cytosine DNA Methylation Is Found in Drosophila Melanogaster but Absent in Saccharomyces Cerevisiae, Schizosaccharomyces Pombe, and Other Yeast Species'. *Analytical Chemistry* 86.8, pp. 3697–3702. DOI: 10.1021/ac500447w.

Carroll, T. M. et al. (2023). 'Tumor Monocyte Content Predicts Immunochemotherapy Outcomes in Esophageal Adenocarcinoma'. *Cancer Cell* 41.7, 1222–1241.e7. DOI: 10.1016/j.ccell.2023.06.006.

Carter, S. L. et al. (2012). 'Absolute Quantification of Somatic DNA Alterations in Human Cancer'. *Nature biotechnology* 30.5, pp. 413–421. DOI: 10.1038/nbt.2203.

Cavalcante, R. G. and M. A. Sartor (2017). 'Annotatr: Genomic Regions in Context'. *Bioinformatics (Oxford, England)* 33.15, pp. 2381–2383. DOI: 10.1093/bioinformatics/btx183.

Chakravarthy, A. et al. (2018). 'Pan-Cancer Deconvolution of Tumour Composition Using DNA Methylation'. *Nature Communications* 9.1 (1), p. 3220. DOI: 10.1038/s41467-018-05570-1.

Chan, H. M. and N. B. La Thangue (2001). 'P300/CBP Proteins: HATs for Transcriptional Bridges and Scaffolds'. *Journal of Cell Science* 114 (Pt 13), pp. 2363–2373. DOI: 10.1242/jcs.114.13.2363.

Chen, J., Q. Deng, Y. Pan, B. He, H. Ying, H. Sun, X. Liu and S. Wang (2015). 'Prognostic Value of Neutrophil-to-Lymphocyte Ratio in Breast Cancer'. *FEBS open bio* 5, pp. 502–507. DOI: 10.1016/j.fob.2015.05.003.

Choy, M.-K., M. Movassagh, H.-G. Goh, M. R. Bennett, T. A. Down and R. S. Y. Foo (2010). 'Genome-Wide Conserved Consensus Transcription Factor Binding Motifs Are Hyper-Methylated'. *BMC genomics* 11, p. 519. DOI: 10.1186/1471-2164-11-519.

Clark, C., P. Palta, C. J. Joyce, C. Scott, E. Grundberg, P. Deloukas, A. Palotie and A. J. Coffey (2012). 'A Comparison of the Whole Genome Approach of MeDIP-Seq to the Targeted Approach of the Infinium HumanMethylation450 BeadChip® for Methylome Profiling'. *PLoS ONE* 7.11, e50233. DOI: 10.1371/journal.pone.0050233.

Clark, S. J., A. Statham, C. Stirzaker, P. L. Molloy and M. Frommer (2006). 'DNA Methylation: Bisulphite Modification and Analysis'. *Nature Protocols* 1.5, pp. 2353–2364. DOI: 10.1038/nprot.2006.324.

Cui, X.-L. et al. (2020). 'A Human Tissue Map of 5-Hydroxymethylcytosines Exhibits Tissue Specificity through Gene and Enhancer Modulation'. *Nature Communications* 11.1 (1), p. 6161. DOI: 10.1038/s41467-020-20001-w.

Cunningham, D. et al. (2008). 'Capecitabine and Oxaliplatin for Advanced Esophagogastric Cancer'. *The New England Journal of Medicine* 358.1, pp. 36–46. DOI: 10.1056/NEJMoa073149.

Dawlaty, M. M. et al. (2013). 'Combined Deficiency of Tet1 and Tet2 Causes Epigenetic Abnormalities but Is Compatible with Postnatal Development'. *Developmental Cell* 24.3, pp. 310–323. DOI: 10.1016/j.devcel.2012.12.015.

De Santa, F., I. Barozzi, F. Mietton, S. Ghisletti, S. Polletti, B. K. Tusi, H. Muller, J. Ragoussis, C.-L. Wei and G. Natoli (2010). 'A Large Fraction of Extragenic RNA Pol II Transcription Sites Overlap Enhancers'. *PLoS biology* 8.5, e1000384. DOI: 10.1371/journal.pbio.1000384.

Deans, C. and K. A. Maggert (2015). 'What Do You Mean, "Epigenetic"?' *Genetics* 199.4, pp. 887–896. DOI: 10.1534/genetics.114.173492.

Devesa, S. S., W. J. Blot and J. F. Fraumeni (1998). 'Changing Patterns in the Incidence of Esophageal and Gastric Carcinoma in the United States'. *Cancer* 83.10, pp. 2049–2053.

Dos Santos Cunha, A. C., A. G. Simon, T. Zander, R. Buettner, C. J. Bruns, W. Schroeder, F. Gebauer and A. Quaas (2023). 'Dissecting the Inflammatory Tumor Microenvironment of Esophageal Adenocarcinoma: Mast Cells and Natural Killer Cells Are Favorable Prognostic Factors and Associated with Less Extensive Disease'. *Journal of Cancer Research and Clinical Oncology* 149.10, pp. 6917–6929. DOI: 10.1007/s00432-023-04650-0.

Dowle, M. and A. Srinivasan (2023). *Data.Table: Extension of 'data.Frame'*. r-datatable.com.

Eckhardt, F. et al. (2006). 'DNA Methylation Profiling of Human Chromosomes 6, 20 and 22'. *Nature Genetics* 38.12, pp. 1378–1385. DOI: 10.1038/ng1909.

Eden, A., F. Gaudet, A. Waghmare and R. Jaenisch (2003). 'Chromosomal Instability and Tumors Promoted by DNA Hypomethylation'. *Science (New York, N.Y.)* 300.5618, p. 455. DOI: 10.1126/science.1083557.

Ehrlich, M. (2009). 'DNA Hypomethylation in Cancer Cells'. *Epigenomics* 1.2, pp. 239–259. DOI: 10.2217/epi.09.33.

ENCODE Project Consortium et al. (2007). 'Identification and Analysis of Functional Elements in 1% of the Human Genome by the ENCODE Pilot Project'. *Nature* 447.7146, pp. 799–816. DOI: 10.1038/nature05874.

Ernst, J. and M. Kellis (2012). 'ChromHMM: Automating Chromatin-State Discovery and Characterization'. *Nature Methods* 9.3, pp. 215–216. DOI: 10.1038/nmeth.1906.

– (2015). 'Large-Scale Imputation of Epigenomic Datasets for Systematic Annotation of Diverse Human Tissues'. *Nature Biotechnology* 33.4, pp. 364–376. DOI: 10.1038/nbt.3157.

Ernst, J. et al. (2011). 'Mapping and Analysis of Chromatin State Dynamics in Nine Human Cell Types'. *Nature* 473.7345 (7345), pp. 43–49. DOI: 10.1038/nature09906.

Esteller, M. (2007). 'Epigenetic Gene Silencing in Cancer: The DNA Hypermethylome'. *Human Molecular Genetics* 16.R1, R50–R59. DOI: 10.1093/hmg/ddm018.

FastQC (2015). URL: https://qubeshub.org/resources/fastqc.

Faustino, N. A. and T. A. Cooper (2003). 'Pre-mRNA Splicing and Human Disease'. *Genes & Development* 17.4, pp. 419–437. DOI: 10.1101/gad.1048803.

FDA (2021a). 'FDA Approves Nivolumab in Combination with Chemotherapy for Metastatic Gastric Cancer and Esophageal Adenocarcinoma'. *FDA*.

– (2021b). 'FDA Approves Pembrolizumab for Esophageal or GEJ Carcinoma'. *FDA*.

Felsenfeld, G. and M. Groudine (2003). 'Controlling the Double Helix'. *Nature* 421.6921, pp. 448–453. DOI: 10.1038/nature01411.

Ferguson-Smith, A. C., H. Sasaki, B. M. Cattanach and M. A. Surani (1993). 'Parental-Origin-Specific Epigenetic Modification of the Mouse H19 Gene'. *Nature* 362.6422 (6422), pp. 751–755. DOI: 10.1038/362751a0.

Ficz, G., M. R. Branco, S. Seisenberger, F. Santos, F. Krueger, T. A. Hore, C. J. Marques, S. Andrews and W. Reik (2011). 'Dynamic Regulation of 5-Hydroxymethylcytosine in Mouse ES Cells and during Differentiation'. *Nature* 473.7347 (7347), pp. 398–402. DOI: 10.1038/nature10008.

Fishilevich, S. et al. (2017). 'GeneHancer: Genome-Wide Integration of Enhancers and Target Genes in GeneCards'. *Database: The Journal of Biological Databases and Curation* 2017, bax028. DOI: 10.1093/database/bax028.

Fitzgerald, R. C. (2004). 'Review Article: Barrett's Oesophagus and Associated Adenocarcinoma–a UK Perspective'. *Alimentary Pharmacology & Therapeutics* 20 Suppl 8, pp. 45–49. DOI: 10.1111/j.1365-2036.2004.02229.x.

Flusberg, B. A., D. Webster, J. Lee, K. Travers, E. Olivares, T. A. Clark, J. Korlach and S. W. Turner (2010). 'Direct Detection of DNA Methylation during Single-Molecule, Real-Time Sequencing'. *Nature methods* 7.6, pp. 461–465. DOI: 10.1038/nmeth.1459.

Frankish, A. et al. (2019). 'GENCODE Reference Annotation for the Human and Mouse Genomes'. *Nucleic Acids Research* 47.D1, pp. D766–D773. DOI: 10.1093/nar/gky955.

Fridman, W. H., F. Pagès, C. Sautès-Fridman and J. Galon (2012). 'The Immune Contexture in Human Tumours: Impact on Clinical Outcome'. *Nature Reviews. Cancer* 12.4, pp. 298–306. DOI: 10.1038/nrc3245.

Gardiner-Garden, M. and M. Frommer (1987). 'CpG Islands in Vertebrate Genomes'. *Journal of Molecular Biology* 196.2, pp. 261–282. DOI: 10.1016/0022-2836(87)90689-9.

Gaudet, F., J. G. Hodgson, A. Eden, L. Jackson-Grusby, J. Dausman, J. W. Gray, H. Leonhardt and R. Jaenisch (2003). 'Induction of Tumors in Mice by Genomic Hypomethylation'. *Science (New York, N.Y.)* 300.5618, pp. 489–492. DOI: 10.1126/science.1083558.

Gaujoux, R. and C. Seoighe (2010). 'A Flexible R Package for Nonnegative Matrix Factorization'. *BMC Bioinformatics* 11.1, p. 367. DOI: 10.1186/1471-2105-11-367.

Gershman, A. et al. (2022). 'Epigenetic Patterns in a Complete Human Genome'. *Science* 376.6588, eabj5089. DOI: 10.1126/science.abj5089.

Ginno, P. A. et al. (2020). 'A Genome-Scale Map of DNA Methylation Turnover Identifies Site-Specific Dependencies of DNMT and TET Activity'. *Nature Communications* 11.1, p. 2680. DOI: 10.1038/s41467-020-16354-x.

Globisch, D., M. Münzel, M. Müller, S. Michalakis, M. Wagner, S. Koch, T. Brückl, M. Biel and T. Carell (2010). 'Tissue Distribution of 5-Hydroxymethylcytosine and Search for Active Demethylation Intermediates'. *PLOS ONE* 5.12, e15367. DOI: 10.1371/journal.pone.0015367.

Greenberg, M. V. C. and D. Bourc'his (2019). 'The Diverse Roles of DNA Methylation in Mammalian Development and Disease'. *Nature Reviews Molecular Cell Biology* 20.10 (10), pp. 590–607. DOI: 10.1038/s41580-019-0159-6.

Gu, T.-P. et al. (2011). 'The Role of Tet3 DNA Dioxygenase in Epigenetic Reprogramming by Oocytes'. *Nature* 477.7366 (7366), pp. 606–610. DOI: 10.1038/nature10443.

Gu, Z., R. Eils and M. Schlesner (2016). 'Complex Heatmaps Reveal Patterns and Correlations in Multidimensional Genomic Data'. *Bioinformatics* 32.18, pp. 2847–2849. DOI: 10.1093/bioinformatics/btw313.

Guibert, S., T. Forné and M. Weber (2012). 'Global Profiling of DNA Methylation Erasure in Mouse Primordial Germ Cells'. *Genome Research* 22.4, pp. 633–641. DOI: 10.1101/gr.130997.111.

Gull, N. et al. (2022). 'DNA Methylation and Transcriptomic Features Are Preserved throughout Disease Recurrence and Chemoresistance in High Grade Serous Ovarian Cancers'. *Journal of experimental & clinical cancer research: CR* 41.1, p. 232. DOI: 10.1186/s13046-022-02440-z.

Guo, J. U., Y. Su, C. Zhong, G.-l. Ming and H. Song (2011). 'Hydroxylation of 5-Methylcytosine by TET1 Promotes Active DNA Demethylation in the Adult Brain'. *Cell* 145.3, p. 423. DOI: 10.1016/j.cell.2011.03.022.

Guo, S., D. Diep, N. Plongthongkum, H.-L. Fung, K. Zhang and K. Zhang (2017). 'Identification of Methylation Haplotype Blocks Aids in Deconvolution of Heterogeneous Tissue Samples and Tumor Tissue-of-Origin Mapping from Plasma DNA'. *Nature Genetics* 49.4, pp. 635–642. DOI: 10.1038/ng.3805.

H, W. C. (1942). 'The Epigenotype'. *Endeavour* 1, pp. 18–20.

Haig, D. (2004). 'The (Dual) Origin of Epigenetics'. *Cold Spring Harbor Symposia on Quantitative Biology* 69, pp. 67–70. DOI: 10.1101/sqb.2004.69.67.

Han, L., I. G. Lin and C.-L. Hsieh (2001). 'Protein Binding Protects Sites on Stable Episomes and in the Chromosome from De Novo Methylation'. *Molecular and Cellular Biology* 21.10, pp. 3416–3424. DOI: 10.1128/MCB.21.10.3416-3424.2001.

He, B. et al. (2021). 'Tissue-Specific 5-Hydroxymethylcytosine Landscape of the Human Genome'. *Nature Communications* 12.1, p. 4249. DOI: 10.1038/s41467-021-24425-w.

He, Y.-F. et al. (2011). 'Tet-Mediated Formation of 5-Carboxylcytosine and Its Excision by TDG in Mammalian DNA'. *Science* 333.6047, pp. 1303–1307. DOI: 10.1126/science.1210944.

He, Y. and J. R. Ecker (2015). 'Non-CG Methylation in the Human Genome'. *Annual Review of Genomics and Human Genetics* 16, pp. 55–77. DOI: 10.1146/annurev-genom-090413-025437.

Heintzman, N. D. et al. (2007). 'Distinct and Predictive Chromatin Signatures of Transcriptional Promoters and Enhancers in the Human Genome'. *Nature Genetics* 39.3 (3), pp. 311–318. DOI: 10.1038/ng1966.

Herman, J. G., F. Latif, Y. Weng, M. I. Lerman, B. Zbar, S. Liu, D. Samid, D. S. Duan, J. R. Gnarra and W. M. Linehan (1994). 'Silencing of the VHL Tumor-Suppressor Gene by DNA Methylation in Renal Carcinoma'. *Proceedings of the National Academy of*

Sciences of the United States of America 91.21, pp. 9700–9704. DOI: 10.1073/pnas.91.21.9700.

Hinrichs, A. S. et al. (2006). 'The UCSC Genome Browser Database: Update 2006'. *Nucleic Acids Research* 34 (Database issue), pp. D590–598. DOI: 10.1093/nar/gkj144.

Hirasawa, R., H. Chiba, M. Kaneda, S. Tajima, E. Li, R. Jaenisch and H. Sasaki (2008). 'Maternal and Zygotic Dnmt1 Are Necessary and Sufficient for the Maintenance of DNA Methylation Imprints during Preimplantation Development'. *Genes & Development* 22.12, pp. 1607–1616. DOI: 10.1101/gad.1667008.

Hnisz, D., B. J. Abraham, T. I. Lee, A. Lau, V. Saint-André, A. A. Sigova, H. A. Hoke and R. A. Young (2013). 'Super-Enhancers in the Control of Cell Identity and Disease'. *Cell* 155.4, pp. 934–947. DOI: 10.1016/j.cell.2013.09.053.

Holliday, R. and J. E. Pugh (1975). 'DNA Modification Mechanisms and Gene Activity during Development'. *Science (New York, N.Y.)* 187.4173, pp. 226–232.

Holliday, R. (1994). 'Epigenetics: An Overview'. *Developmental Genetics* 15.6, pp. 453–457. DOI: 10.1002/dvg.1020150602.

Holm, T. M., L. Jackson-Grusby, T. Brambrink, Y. Yamada, W. M. Rideout and R. Jaenisch (2005). 'Global Loss of Imprinting Leads to Widespread Tumorigenesis in Adult Mice'. *Cancer Cell* 8.4, pp. 275–285. DOI: 10.1016/j.ccr.2005.09.007.

Homo Sapiens Genome Assembly T2T-CHM13v2.0 (2023). NCBI. URL: https://www.ncbi.nlm.nih.gov/data-hub/assembly/GCF_009914755.1/ (visited on 11/08/2023).

Hon, G. C. et al. (2014). '5mC Oxidation by Tet2 Modulates Enhancer Activity and Timing of Transcriptome Reprogramming during Differentiation'. *Molecular Cell* 56.2, pp. 286–297. DOI: 10.1016/j.molcel.2014.08.026.

Houseman, E. A., M. L. Kile, D. C. Christiani, T. A. Ince, K. T. Kelsey and C. J. Marsit (2016). 'Reference-Free Deconvolution of DNA Methylation Data and Mediation by Cell Composition Effects'. *BMC Bioinformatics* 17.1, p. 259. DOI: 10.1186/s12859-016-1140-4.

Hsieh, C. L. (1994). 'Dependence of Transcriptional Repression on CpG Methylation Density.' *Molecular and Cellular Biology* 14.8, pp. 5487–5494.

Iguchi-Ariga, S. M. and W. Schaffner (1989). 'CpG Methylation of the cAMP-responsive Enhancer/Promoter Sequence TGACGTCA Abolishes Specific Factor Binding as Well as Transcriptional Activation.' *Genes & Development* 3.5, pp. 612–619. DOI: 10.1101/gad.3.5.612.

Illingworth, R. S. and A. P. Bird (2009). 'CpG Islands–'a Rough Guide". *FEBS letters* 583.11, pp. 1713–1720. DOI: 10.1016/j.febslet.2009.04.012.

Illingworth, R. S., U. Gruenewald-Schneider, S. Webb, A. R. W. Kerr, K. D. James, D. J. Turner, C. Smith, D. J. Harrison, R. Andrews and A. P. Bird (2010). 'Orphan CpG Islands Identify Numerous Conserved Promoters in the Mammalian Genome'. *PLOS Genetics* 6.9, e1001134. DOI: 10.1371/journal.pgen.1001134.

Irizarry, R. A. et al. (2009). 'The Human Colon Cancer Methylome Shows Similar Hypo- and Hypermethylation at Conserved Tissue-Specific CpG Island Shores'. *Nature Genetics* 41.2, pp. 178–186. DOI: 10.1038/ng.298.

Ishibashi, S., Y. Ohashi, T. Suzuki, S. Miyazaki, T. Moriya, S. Satomi and H. Sasano (2006). 'Tumor-Associated Tissue Eosinophilia in Human Esophageal Squamous Cell Carcinoma'. *Anticancer Research* 26 (2B), pp. 1419–1424.

Issa, J.-P. (2004). 'CpG Island Methylator Phenotype in Cancer'. *Nature Reviews Cancer* 4.12 (12), pp. 988–993. DOI: 10.1038/nrc1507.

Ito, S., L. Shen, Q. Dai, S. C. Wu, L. B. Collins, J. A. Swenberg, C. He and Y. Zhang (2011). 'Tet Proteins Can Convert 5-Methylcytosine to 5-Formylcytosine and 5-Carboxylcytosine'. *Science (New York, N.Y.)* 333.6047, pp. 1300–1303. DOI: 10.1126/science.1210597.

Ivanov, M., M. Kals, M. Kacevska, I. Barragan, K. Kasuga, A. Rane, A. Metspalu, L. Milani and M. Ingelman-Sundberg (2013). 'Ontogeny, Distribution and Potential Roles of 5-Hydroxymethylcytosine in Human Liver Function'. *Genome Biology* 14.8, R83. DOI: 10.1186/gb-2013-14-8-r83.

Jablonka, E. and M. J. Lamb (2002). 'The Changing Concept of Epigenetics'. *Annals of the New York Academy of Sciences* 981, pp. 82–96. DOI: 10.1111/j.1749-6632.2002.tb04913.x.

Jackson, K., M. C. Yu, K. Arakawa, E. Fiala, B. Youn, H. Fiegl, E. Müller-Holzner, M. Widschwendter and M. Ehrlich (2004). 'DNA Hypomethylation Is Prevalent Even in Low-Grade Breast Cancers'. *Cancer Biology & Therapy* 3.12, pp. 1225–1231. DOI: 10.4161/cbt.3.12.1222.

Jaenisch, R. and A. Bird (2003). 'Epigenetic Regulation of Gene Expression: How the Genome Integrates Intrinsic and Environmental Signals'. *Nature Genetics* 33 Suppl, pp. 245–254. DOI: 10.1038/ng1089.

Jaffe, A. E. and R. A. Irizarry (2014). 'Accounting for Cellular Heterogeneity Is Critical in Epigenome-Wide Association Studies'. *Genome Biology* 15.2, R31. DOI: 10.1186/gb-2014-15-2-r31.

Jatoi, A., B. R. Murphy, N. R. Foster, D. A. Nikcevich, S. R. Alberts, J. A. Knost, T. R. Fitch and K. M. Rowland (2006). 'Oxaliplatin and Capecitabine in Patients with Metastatic Adenocarcinoma of the Esophagus, Gastroesophageal Junction and Gastric Cardia: A Phase II Study from the North Central Cancer Treatment Group'. *Annals of Oncology* 17.1, pp. 29–34. DOI: 10.1093/annonc/mdj063.

Jeziorska, D. M. et al. (2017). 'DNA Methylation of Intragenic CpG Islands Depends on Their Transcriptional Activity during Differentiation and Disease'. *Proceedings of the National Academy of Sciences* 114.36, E7526–E7535. DOI: 10.1073/pnas.1703087114.

Jiao, Y., M. Widschwendter and A. E. Teschendorff (2014). 'A Systems-Level Integrative Framework for Genome-Wide DNA Methylation and Gene Expression Data Identifies Differential Gene Expression Modules under Epigenetic Control'. *Bioinformatics (Oxford, England)* 30.16, pp. 2360–2366. DOI: 10.1093/bioinformatics/btu316.

Jin, M.-Z. and W.-L. Jin (2020). 'The Updated Landscape of Tumor Microenvironment and Drug Repurposing'. *Signal Transduction and Targeted Therapy* 5.1 (1), pp. 1–16. DOI: 10.1038/s41392-020-00280-x.

Jin, S.-G., Y. Jiang, R. Qiu, T. A. Rauch, Y. Wang, G. Schackert, D. Krex, Q. Lu and G. P. Pfeifer (2011a). '5-Hydroxymethylcytosine Is Strongly Depleted in Human Cancers but Its Levels Do Not Correlate with IDH1 Mutations'. *Cancer research* 71.24, pp. 7360–7365. DOI: 10.1158/0008-5472.CAN-11-2023.

Jin, S.-G., X. Wu, A. X. Li and G. P. Pfeifer (2011b). 'Genomic Mapping of 5-Hydroxymethylcytosine in the Human Brain'. *Nucleic Acids Research* 39.12, pp. 5015–5024. DOI: 10.1093/nar/gkr120.

Johnson, K. C., E. A. Houseman, J. E. King, K. M. von Herrmann, C. E. Fadul and B. C. Christensen (2016). '5-Hydroxymethylcytosine Localizes to Enhancer Elements and Is Associated with Survival in Glioblastoma Patients'. *Nature Communications* 7.1 (1), p. 13177. DOI: 10.1038/ncomms13177.

Jones, P. A. (1999). 'The DNA Methylation Paradox'. *Trends in genetics: TIG* 15.1, pp. 34–37. DOI: 10.1016/s0168-9525(98)01636-9.

Jones, P. A. and D. Takai (2001). 'The Role of DNA Methylation in Mammalian Epigenetics'. *Science (New York, N.Y.)* 293.5532, pp. 1068–1070. DOI: 10.1126/science.1063852.

Jones, P. A. (2012). 'Functions of DNA Methylation: Islands, Start Sites, Gene Bodies and Beyond'. *Nature Reviews Genetics* 13.7 (7), pp. 484–492. DOI: 10.1038/nrg3230.

Juven-Gershon, T., J.-Y. Hsu, J. W. Theisen and J. T. Kadonaga (2008). 'The RNA Polymerase II Core Promoter - the Gateway to Transcription'. *Current Opinion in Cell Biology* 20.3, pp. 253–259. DOI: 10.1016/j.ceb.2008.03.003.

Karlsson, M. et al. (2021). 'A Single–Cell Type Transcriptomics Map of Human Tissues'. *Science Advances* 7.31, eabh2169. DOI: 10.1126/sciadv.abh2169.

Khatri, P., M. Sirota and A. J. Butte (2012). 'Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges'. *PLOS Computational Biology* 8.2, e1002375. DOI: 10.1371/journal.pcbi.1002375.

Kim, J. et al. (2017). 'Integrated Genomic Characterization of Oesophageal Carcinoma'. *Nature* 541.7636 (7636), pp. 169–175. DOI: 10.1038/nature20805.

Kim, T.-K. et al. (2010). 'Widespread Transcription at Neuronal Activity-Regulated Enhancers'. *Nature* 465.7295, pp. 182–187. DOI: 10.1038/nature09033.

Kitts, A. and S. Sherry (2011). 'The Single Nucleotide Polymorphism Database (dbSNP) of Nucleotide Sequence Variation'. *The NCBI Handbook [Internet]*. National Center for Biotechnology Information (US).

Klein, C. J., T. Bird, N. Ertekin-Taner, S. Lincoln, R. Hjorth, Y. Wu, J. Kwok, G. Mer, P. J. Dyck and G. A. Nicholson (2013). 'DNMT1 Mutation Hot Spot Causes Varied Phenotypes of HSAN1 with Dementia and Hearing Loss'. *Neurology* 80.9, pp. 824–828. DOI: 10.1212/WNL.0b013e318284076d.

Klutstein, M., D. Nejman, R. Greenfield and H. Cedar (2016). 'DNA Methylation in Cancer and Aging'. *Cancer Research* 76.12, pp. 3446–3450. DOI: 10.1158/0008-5472.CAN-15-3278.

Koestler, D. C., M. J. Jones, J. Usset, B. C. Christensen, R. A. Butler, M. S. Kobor, J. K. Wiencke and K. T. Kelsey (2016). 'Improving Cell Mixture Deconvolution by Identifying Optimal DNA Methylation Libraries (IDOL)'. *BMC Bioinformatics* 17.1, p. 120. DOI: 10.1186/s12859-016-0943-7.

Kress, C., H. Thomassin and T. Grange (2006). 'Active Cytosine Demethylation Triggered by a Nuclear Receptor Involves DNA Strand Breaks'. *Proceedings of the National Academy of Sciences* 103.30, pp. 11112–11117. DOI: 10.1073/pnas.0601793103.

Kriaucionis, S. and N. Heintz (2009). 'The Nuclear DNA Base 5-Hydroxymethylcytosine Is Present in Purkinje Neurons and the Brain'. *Science (New York, N.Y.)* 324.5929, pp. 929–930. DOI: 10.1126/science.1169786.

Krueger, F. (2023). *Trim Galore*. URL: https://github.com/FelixKrueger/TrimGalore (visited on 03/08/2023).

Ku, M. et al. (2008). 'Genomewide Analysis of PRC1 and PRC2 Occupancy Identifies Two Classes of Bivalent Domains'. *PLOS Genetics* 4.10, e1000242. DOI: 10.1371/journal.pgen.1000242.

Kudo, Y., K. Tateishi, K. Yamamoto, S. Yamamoto, Y. Asaoka, H. Ijichi, G. Nagae, H. Yoshida, H. Aburatani and K. Koike (2012). 'Loss of 5-Hydroxymethylcytosine Is Accompanied with Malignant Cellular Transformation'. *Cancer Science* 103.4, pp. 670–676. DOI: 10.1111/j.1349-7006.2012.02213.x.

Kundaje, A. et al. (2015). 'Integrative Analysis of 111 Reference Human Epigenomes'. *Nature* 518.7539 (7539), pp. 317–330. DOI: 10.1038/nature14248.

Landt, S. G. et al. (2012). 'ChIP-seq Guidelines and Practices of the ENCODE and modENCODE Consortia'. *Genome Research* 22.9, pp. 1813–1831. DOI: 10.1101/gr.136184.111.

Larsen, F., G. Gundersen, R. Lopez and H. Prydz (1992). 'CpG Islands as Gene Markers in the Human Genome'. *Genomics* 13.4, pp. 1095–1107. DOI: 10.1016/0888-7543(92)90024-m.

Laurent, L. et al. (2010). 'Dynamic Changes in the Human Methylome during Differentiation'. *Genome Research* 20.3, pp. 320–331. DOI: 10.1101/gr.101907.109.

Law, J. A. and S. E. Jacobsen (2010). 'Establishing, Maintaining and Modifying DNA Methylation Patterns in Plants and Animals'. *Nature Reviews Genetics* 11.3 (3), pp. 204–220. DOI: 10.1038/nrg2719.

Lawrence, M., W. Huber, H. Pagès, P. Aboyoun, M. Carlson, R. Gentleman, M. T. Morgan and V. J. Carey (2013). 'Software for Computing and Annotating Genomic Ranges'. *PLOS Computational Biology* 9.8, e1003118. DOI: 10.1371/journal.pcbi.1003118.

Lee, D. D. and H. S. Seung (1999). 'Learning the Parts of Objects by Non-Negative Matrix Factorization'. *Nature* 401.6755 (6755), pp. 788–791. DOI: 10.1038/44565.

Lee, S. and S. Ditko (1962). *Amazing Fantasy*. Vol. 15. Marvel Comics.

Leers, J. M., S. R. DeMeester, N. Chan, S. Ayazi, A. Oezcelik, E. Abate, F. Banki, J. C. Lipham, J. A. Hagen and T. R. DeMeester (2009). 'Clinical Characteristics, Biologic Behavior, and Survival after Esophagectomy Are Similar for Adenocarcinoma of the Gastroesophageal Junction and the Distal Esophagus'. *The Journal of Thoracic and Cardiovascular Surgery* 138.3, pp. 594–602. DOI: 10.1016/j.jtcvs.2009.05.039.

Lettice, L. A., S. J. Heaney, L. A. Purdie, L. Li, P. de Beer, B. A. Oostra, D. Goode, G. Elgar, R. E. Hill and E. de Graaff (2003). 'A Long-Range Shh Enhancer Regulates Expression in the Developing Limb and Fin and Is Associated with Preaxial Polydactyly'. *Human Molecular Genetics* 12.14, pp. 1725–1735. DOI: 10.1093/hmg/ddg180.

Levy, S. E. and R. M. Myers (2016). 'Advancements in Next-Generation Sequencing'. *Annual Review of Genomics and Human Genetics* 17.1, pp. 95–115. DOI: 10.1146/annurev-genom-083115-022413.

Li, E., T. H. Bestor and R. Jaenisch (1992). 'Targeted Mutation of the DNA Methyltransferase Gene Results in Embryonic Lethality'. *Cell* 69.6, pp. 915–926. DOI: 10.1016/0092-8674(92)90611-f.

Li, E., C. Beard and R. Jaenisch (1993). 'Role for DNA Methylation in Genomic Imprinting'. *Nature* 366.6453 (6453), pp. 362–365. DOI: 10.1038/366362a0.

Lian, C. G. et al. (2012). 'Loss of 5-Hydroxymethylcytosine Is an Epigenetic Hallmark of Melanoma'. *Cell* 150.6, pp. 1135–1146. DOI: 10.1016/j.cell.2012.07.033.

Liberzon, A., A. Subramanian, R. Pinchback, H. Thorvaldsdóttir, P. Tamayo and J. P. Mesirov (2011). 'Molecular Signatures Database (MSigDB) 3.0'. *Bioinformatics* 27.12, pp. 1739–1740. DOI: 10.1093/bioinformatics/btr260.

Lindahl, T. and B. Nyberg (1974). 'Heat-Induced Deamination of Cytosine Residues in Deoxyribonucleic Acid'. *Biochemistry* 13.16, pp. 3405–3410. DOI: 10.1021/bi00713a035.

Lister, R. et al. (2009). 'Human DNA Methylomes at Base Resolution Show Widespread Epigenomic Differences'. *Nature* 462.7271, pp. 315–322. DOI: 10.1038/nature08514.

Litchfield, K. et al. (2021). 'Meta-Analysis of Tumor- and T Cell-Intrinsic Mechanisms of Sensitization to Checkpoint Inhibition'. *Cell* 184.3, 596–614.e14. DOI: 10.1016/j.cell.2021.01.002.

Liu, Y., Z. Hu, J. Cheng, P. Siejka-Zielińska, J. Chen, M. Inoue, A. A. Ahmed and C.-X. Song (2021). 'Subtraction-Free and Bisulfite-Free Specific Sequencing of 5-Methylcytosine and Its Oxidized Derivatives at Base Resolution'. *Nature Communications* 12.1 (1), p. 618. DOI: 10.1038/s41467-021-20920-2.

Liu, Y., P. Siejka-Zielińska, G. Velikova, Y. Bi, F. Yuan, M. Tomkova, C. Bai, L. Chen, B. Schuster-Böckler and C.-X. Song (2019). 'Bisulfite-Free Direct Detection of 5-Methylcytosine and 5-Hydroxymethylcytosine at Base Resolution'. *Nature Biotechnology* 37.4 (4), pp. 424–429. DOI: 10.1038/s41587-019-0041-2.

Locke, W. J., D. Guanzon, C. Ma, Y. J. Liew, K. R. Duesing, K. Y. Fung and J. P. Ross (2019). 'DNA Methylation Cancer Biomarkers: Translation to the Clinic'. *Frontiers in Genetics* 10.

Loyfer, N. et al. (2023). 'A DNA Methylation Atlas of Normal Human Cell Types'. *Nature* 613.7943 (7943), pp. 355–364. DOI: 10.1038/s41586-022-05580-6.

Lu, F., Y. Liu, L. Jiang, S. Yamaguchi and Y. Zhang (2014). 'Role of Tet Proteins in Enhancer Activity and Telomere Elongation'. *Genes & Development* 28.19, pp. 2103–2119. DOI: 10.1101/gad.248005.114.

Maiti, A. and A. C. Drohat (2011). 'Thymine DNA Glycosylase Can Rapidly Excise 5-Formylcytosine and 5-Carboxylcytosine: Potential Implications for Active Demethylation of CpG Sites'. *The Journal of Biological Chemistry* 286.41, pp. 35334–35338. DOI: 10.1074/jbc.C111.284620.

Margueron, R. and D. Reinberg (2010). 'Chromatin Structure and the Inheritance of Epigenetic Information'. *Nature Reviews. Genetics* 11.4, pp. 285–296. DOI: 10.1038/nrg2752.

Maunakea, A. K. et al. (2010). 'Conserved Role of Intragenic DNA Methylation in Regulating Alternative Promoters'. *Nature* 466.7303, pp. 253–257. DOI: 10.1038/nature09165.

McGinty, R. K. and S. Tan (2015). 'Nucleosome Structure and Function'. *Chemical Reviews* 115.6, pp. 2255–2273. DOI: 10.1021/cr500373h.

Meissner, A. et al. (2008). 'Genome-Scale DNA Methylation Maps of Pluripotent and Differentiated Cells'. *Nature* 454.7205, pp. 766–770. DOI: 10.1038/nature07107.

Mikkelsen, T. S. et al. (2007). 'Genome-Wide Maps of Chromatin State in Pluripotent and Lineage-Committed Cells'. *Nature* 448.7153 (7153), pp. 553–560. DOI: 10.1038/nature06008.

Moore, L. D., T. Le and G. Fan (2013). 'DNA Methylation and Its Basic Function'. *Neuropsychopharmacology* 38.1 (1), pp. 23–38. DOI: 10.1038/npp.2012.112.

Morales, J. et al. (2022). 'A Joint NCBI and EMBL-EBI Transcript Set for Clinical Genomics and Research'. *Nature* 604.7905, pp. 310–315. DOI: 10.1038/s41586-022-04558-8.

Moss, J. et al. (2018). 'Comprehensive Human Cell-Type Methylation Atlas Reveals Origins of Circulating Cell-Free DNA in Health and Disease'. *Nature Communications* 9.1 (1), pp. 1–12. DOI: 10.1038/s41467-018-07466-6.

Nassar, L. R. et al. (2023). 'The UCSC Genome Browser Database: 2023 Update'. *Nucleic Acids Research* 51.D1, pp. D1188–D1195. DOI: 10.1093/nar/gkac1072.

Neri, F., S. Rapelli, A. Krepelova, D. Incarnato, C. Parlato, G. Basile, M. Maldotti, F. Anselmi and S. Oliviero (2017). 'Intragenic DNA Methylation Prevents Spurious Transcription Initiation'. *Nature* 543.7643, pp. 72–77. DOI: 10.1038/nature21373.

Nestor, C. E., R. Ottaviano, J. Reddington, D. Sproul, D. Reinhardt, D. Dunican, E. Katz, J. M. Dixon, D. J. Harrison and R. R. Meehan (2012). 'Tissue Type Is a Major Modifier of the 5-Hydroxymethylcytosine Content of Human Genes'. *Genome Research* 22.3, pp. 467–477. DOI: 10.1101/gr.126417.111.

Nik-Zainal, S. et al. (2012). 'Mutational Processes Molding the Genomes of 21 Breast Cancers'. *Cell* 149.5, pp. 979–993. DOI: 10.1016/j.cell.2012.04.024.

Nowicki-Osuch, K. et al. (2021). 'Molecular Phenotyping Reveals the Identity of Barrett's Esophagus and Its Malignant Transition'. *Science* 373.6556, pp. 760–767. DOI: 10.1126/science.abd1449.

Nurk, S. et al. (2022). 'The Complete Sequence of a Human Genome'. *Science* 376.6588, pp. 44–53. DOI: 10.1126/science.abj6987.

Ohashi, Y., S. Ishibashi, T. Suzuki, R. Shineha, T. Moriya, S. Satomi and H. Sasano (2000). 'Significance of Tumor Associated Tissue Eosinophilia and Other Inflammatory Cell Infiltrate in Early Esophageal Squamous Cell Carcinoma'. *Anticancer Research* 20 (5A), pp. 3025–3030.

Okano, M., D. W. Bell, D. A. Haber and E. Li (1999). 'DNA Methyltransferases Dnmt3a and Dnmt3b Are Essential for de Novo Methylation and Mammalian Development'. *Cell* 99.3, pp. 247–257. DOI: 10.1016/s0092-8674(00)81656-6.

Ooi, S. K. T. et al. (2007). 'DNMT3L Connects Unmethylated Lysine 4 of Histone H3 to de Novo Methylation of DNA'. *Nature* 448.7154 (7154), pp. 714–717. DOI: 10.1038/nature05987.

Pastor, W. A. et al. (2011). 'Genome-Wide Mapping of 5-Hydroxymethylcytosine in Embryonic Stem Cells'. *Nature* 473.7347, pp. 394–397. DOI: 10.1038/nature10102.

Pennathur, A., M. K. Gibson, B. A. Jobe and J. D. Luketich (2013). 'Oesophageal Carcinoma'. *Lancet (London, England)* 381.9864, pp. 400–412. DOI: 10.1016/S0140-6736(12)60643-6.

Peters, Y. et al. (2019). 'Barrett Oesophagus'. *Nature Reviews. Disease Primers* 5.1, p. 35. DOI: 10.1038/s41572-019-0086-z.

Picard Tools - By Broad Institute (2023). URL: https://broadinstitute.github.io/picard/ (visited on 03/08/2023).

Plass, C. and P. D. Soloway (2002). 'DNA Methylation, Imprinting and Cancer'. *European journal of human genetics: EJHG* 10.1, pp. 6–16. DOI: 10.1038/sj.ejhg.5200768.

Pomraning, K. R., K. M. Smith and M. Freitag (2009). 'Genome-Wide High Throughput Analysis of DNA Methylation in Eukaryotes'. *Methods (San Diego, Calif.)* 47.3, pp. 142–150. DOI: 10.1016/j.ymeth.2008.09.022.

Quante, M. et al. (2012). 'Bile Acid and Inflammation Activate Gastric Cardia Stem Cells in a Mouse Model of Barrett-Like Metaplasia'. *Cancer Cell* 21.1, pp. 36–51. DOI: 10.1016/j.ccr.2011.12.004.

Quinlan, A. R. and I. M. Hall (2010). 'BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features'. *Bioinformatics* 26.6, pp. 841–842. DOI: 10.1093/bioinformatics/btq033.

Quivoron, C. et al. (2011). 'TET2 Inactivation Results in Pleiotropic Hematopoietic Abnormalities in Mouse and Is a Recurrent Event during Human Lymphomagenesis'. *Cancer Cell* 20.1, pp. 25–38. DOI: 10.1016/j.ccr.2011.06.003.

Rahbari, R. et al. (2016). 'Timing, Rates and Spectra of Human Germline Mutation'. *Nature Genetics* 48.2 (2), pp. 126–133. DOI: 10.1038/ng.3469.

Raiber, E.-A., R. Hardisty, P. van Delft and S. Balasubramanian (2017). 'Mapping and Elucidating the Function of Modified Bases in DNA'. *Nature Reviews Chemistry* 1.9 (9), pp. 1–13. DOI: 10.1038/s41570-017-0069.

Raine, K. M., P. Van Loo, D. C. Wedge, D. Jones, A. Menzies, A. P. Butler, J. W. Teague, P. Tarpey, S. Nik-Zainal and P. J. Campbell (2016). 'ascatNgs: Identifying Somatically Acquired Copy-Number Alterations from Whole-Genome Sequencing Data'. *Current protocols in bioinformatics* 56, pp. 15.9.1–15.9.17. DOI: 10.1002/cpbi.17.

Rauch, T. A., X. Wu, X. Zhong, A. D. Riggs and G. P. Pfeifer (2009). 'A Human B Cell Methylome at 100base Pair Resolution'. *Proceedings of the National Academy of Sciences* 106.3, pp. 671–678. DOI: 10.1073/pnas.0812399106.

Reynolds, G. A., S. K. Basu, T. F. Osborne, D. J. Chin, G. Gil, M. S. Brown, J. L. Goldstein and K. L. Luskey (1984). 'HMG CoA Reductase: A Negatively Regulated Gene with

Unusual Promoter and 5' Untranslated Regions'. *Cell* 38.1, pp. 275–285. DOI: 10.1016/0092-8674(84)90549-x.

Rideout, W. M., G. A. Coetzee, A. F. Olumi and P. A. Jones (1990). '5-Methylcytosine as an Endogenous Mutagen in the Human LDL Receptor and P53 Genes'. *Science (New York, N.Y.)* 249.4974, pp. 1288–1290. DOI: 10.1126/science.1697983.

Riggs, A. D. (1975). 'X Inactivation, Differentiation, and DNA Methylation'. *Cytogenetics and Cell Genetics* 14.1, pp. 9–25. DOI: 10.1159/000130315.

Robert, C. (2020). 'A Decade of Immune-Checkpoint Inhibitors in Cancer Therapy'. *Nature Communications* 11.1, p. 3801. DOI: 10.1038/s41467-020-17670-y.

Ryan, D. (2023). *MethylDackel*. URL: https://github.com/dpryan79/MethylDackel (visited on 03/08/2023).

Saghafinia, S., M. Mina, N. Riggi, D. Hanahan and G. Ciriello (2018). 'Pan-Cancer Landscape of Aberrant DNA Methylation across Human Tumors'. *Cell Reports* 25.4, 1066–1080.e8. DOI: 10.1016/j.celrep.2018.09.082.

Saxonov, S., P. Berg and D. L. Brutlag (2006). 'A Genome-Wide Analysis of CpG Dinucleotides in the Human Genome Distinguishes Two Distinct Classes of Promoters'. *Proceedings of the National Academy of Sciences of the United States of America* 103.5, pp. 1412–1417. DOI: 10.1073/pnas.0510310103.

Schmidl, C., M. Klug, T. J. Boeld, R. Andreesen, P. Hoffmann, M. Edinger and M. Rehli (2009). 'Lineage-Specific DNA Methylation in T Cells Correlates with Histone Methylation and Enhancer Activity'. *Genome Research* 19.7, pp. 1165–1174. DOI: 10.1101/gr.091470.109.

Schmutte, C., A. S. Yang, T. T. Nguyen, R. W. Beart and P. A. Jones (1996). 'Mechanisms for the Involvement of DNA Methylation in Colon Carcinogenesis'. *Cancer Research* 56.10, pp. 2375–2381.

Ségurel, L., M. J. Wyman and M. Przeworski (2014). 'Determinants of Mutation Rate Variation in the Human Germline'. *Annual Review of Genomics and Human Genetics* 15.1, pp. 47–70. DOI: 10.1146/annurev-genom-031714-125740.

Sen, S. K., K. Han, J. Wang, J. Lee, H. Wang, P. A. Callinan, M. Dyer, R. Cordaux, P. Liang and M. A. Batzer (2006). 'Human Genomic Deletions Mediated by Recombination between Alu Elements'. *American Journal of Human Genetics* 79.1, pp. 41–53. DOI: 10.1086/504600.

Shapiro, J. A. and R. von Sternberg (2005). 'Why Repetitive DNA Is Essential to Genome Function'. *Biological Reviews* 80.2, pp. 227–250. DOI: 10.1017/S1464793104006657.

Shapiro, R., B. Braverman, J. B. Louis and R. E. Servis (1973). 'Nucleic Acid Reactivity and Conformation. II. Reaction of Cytosine and Uracil with Sodium Bisulfite'. *The Journal of Biological Chemistry* 248.11, pp. 4060–4064.

Siewert, J. R. and K. Ott (2007). 'Are Squamous and Adenocarcinomas of the Esophagus the Same Disease?' *Seminars in Radiation Oncology*. Esophageal Cancer 17.1, pp. 38–44. DOI: 10.1016/j.semradonc.2006.09.007.

Simpson, V. J., T. E. Johnson and R. F. Hammen (1986). 'Caenorhabditis Elegans DNA Does Not Contain 5-Methylcytosine at Any Time during Development or Aging.' *Nucleic Acids Research* 14.16, pp. 6711–6719.

Sims, D., I. Sudbery, N. E. Ilott, A. Heger and C. P. Ponting (2014). 'Sequencing Depth and Coverage: Key Considerations in Genomic Analyses'. *Nature Reviews Genetics* 15.2 (2), pp. 121–132. DOI: 10.1038/nrg3642.

Smyth, E. C., J. Lagergren, R. C. Fitzgerald, F. Lordick, M. A. Shah, P. Lagergren and D. Cunningham (2017). 'Oesophageal Cancer'. *Nature Reviews. Disease Primers* 3, p. 17048. DOI: 10.1038/nrdp.2017.48.

Song, C.-X., C. Yi and C. He (2012). 'Mapping New Nucleotide Variants in the Genome and Transcriptome'. *Nature biotechnology* 30.11, pp. 1107–1116. DOI: 10.1038/nbt.2398.

Song, C.-X. et al. (2011). 'Selective Chemical Labeling Reveals the Genome-Wide Distribution of 5-Hydroxymethylcytosine'. *Nature Biotechnology* 29.1 (1), pp. 68–72. DOI: 10.1038/nbt.1732.

Song, F., J. F. Smith, M. T. Kimura, A. D. Morrow, T. Matsuyama, H. Nagase and W. A. Held (2005). 'Association of Tissue-Specific Differentially Methylated Regions (TDMs) with Differential Gene Expression'. *Proceedings of the National Academy of Sciences* 102.9, pp. 3336–3341. DOI: 10.1073/pnas.0408436102.

Stadler, M. B. et al. (2011). 'DNA-binding Factors Shape the Mouse Methylome at Distal Regulatory Regions'. *Nature* 480.7378, pp. 490–495. DOI: 10.1038/nature10716.

Stirzaker, C., P. C. Taberlay, A. L. Statham and S. J. Clark (2014). 'Mining Cancer Methylomes: Prospects and Challenges'. *Trends in Genetics* 30.2, pp. 75–84. DOI: 10.1016/j.tig.2013.11.004.

Stroud, H., S. Feng, S. Morey Kinney, S. Pradhan and S. E. Jacobsen (2011). '5-Hydroxymethylcytosine Is Associated with Enhancers and Gene Bodies in Human Embryonic Stem Cells'. *Genome Biology* 12.6, R54. DOI: 10.1186/gb-2011-12-6-r54.

Strunnikova, M., U. Schagdarsurengin, A. Kehlen, J. C. Garbe, M. R. Stampfer and R. Dammann (2005). 'Chromatin Inactivation Precedes de Novo DNA Methylation during the Progressive Epigenetic Silencing of the RASSF1A Promoter'. *Molecular and Cellular Biology* 25.10, pp. 3923–3933. DOI: 10.1128/MCB.25.10.3923-3933.2005.

Su, Z., L. Han and Z. Zhao (2011). 'Conservation and Divergence of DNA Methylation in Eukaryotes'. *Epigenetics* 6.2, pp. 134–140. DOI: 10.4161/epi.6.2.13875.

Suh, Y.-S., D.-S. Han, S.-H. Kong, H.-J. Lee, Y. T. Kim, W.-H. Kim, K. U. Lee and H.-K. Yang (2012). 'Should Adenocarcinoma of the Esophagogastric Junction Be Classified as

Esophageal Cancer? A Comparative Analysis According to the Seventh AJCC TNM Classification'. *Annals of Surgery* 255.5, p. 908. DOI: 10.1097/SLA.0b013e31824beb95.

Supek, F., B. Lehner, P. Hajkova and T. Warnecke (2014). 'Hydroxymethylated Cytosines Are Associated with Elevated C to G Transversion Rates'. *PLoS Genetics* 10.9. Ed. by L. Duret, e1004585. DOI: 10.1371/journal.pgen.1004585.

Suzuki, K., I. Suzuki, A. Leodolter, S. Alonso, S. Horiuchi, K. Yamashita and M. Perucho (2006). 'Global DNA Demethylation in Gastrointestinal Cancer Is Age Dependent and Precedes Genomic Damage'. *Cancer Cell* 9.3, pp. 199–207. DOI: 10.1016/j.ccr.2006.02.016.

Suzuki, M. M. and A. Bird (2008). 'DNA Methylation Landscapes: Provocative Insights from Epigenomics'. *Nature Reviews. Genetics* 9.6, pp. 465–476. DOI: 10.1038/nrg2341.

Swisher, E. M., M. I. Harrell, K. Lin, R. L. Coleman, G. E. Konecny, A. V. Tinker, D. M. O'Malley, I. McNeish and S. H. Kaufmann (2017). 'BRCA1 and RAD51C Promoter Hypermethylation Confer Sensitivity to the PARP Inhibitor Rucaparib in Patients with Relapsed, Platinum-Sensitive Ovarian Carcinoma in ARIEL2 Part 1'. *Gynecologic Oncology* 145, p. 5. DOI: 10.1016/j.ygyno.2017.03.034.

Syed, A., C. Maradey-Romero and R. Fass (2017). 'The Relationship between Eosinophilic Esophagitis and Esophageal Cancer'. *Diseases of the Esophagus* 30.7, pp. 1–5. DOI: 10.1093/dote/dox050.

Szulwach, K. E. et al. (2011). '5-hmC-mediated Epigenetic Dynamics during Postnatal Neurodevelopment and Aging'. *Nature Neuroscience* 14.12, pp. 1607–1616. DOI: 10.1038/nn.2959.

Tahiliani, M. et al. (2009). 'Conversion of 5-Methylcytosine to 5-Hydroxymethylcytosine in Mammalian DNA by MLL Partner TET1'. *Science (New York, N.Y.)* 324.5929, pp. 930–935. DOI: 10.1126/science.1170116.

Takai, D. and P. A. Jones (2002). 'Comprehensive Analysis of CpG Islands in Human Chromosomes 21 and 22'. *Proceedings of the National Academy of Sciences of the United States of America* 99.6, pp. 3740–3745. DOI: 10.1073/pnas.052410099.

Tate, P. H. and A. P. Bird (1993). 'Effects of DNA Methylation on DNA-binding Proteins and Gene Expression'. *Current Opinion in Genetics & Development* 3.2, pp. 226–231. DOI: 10.1016/0959-437x(93)90027-m.

Team TBD (2023). *BSgenome.Hsapiens.UCSC.Hg38: Full Genome Sequences for Homo Sapiens (UCSC Version Hg38, Based on GRCh38.P13)*. R package version 1.4.5.

The ENCODE Project Consortium (2012). 'An Integrated Encyclopedia of DNA Elements in the Human Genome'. *Nature* 489.7414, pp. 57–74. DOI: 10.1038/nature11247.

The Human Protein Atlas (2023). URL: https://www.proteinatlas.org/ (visited on 02/08/2023).

Thomson, J. P. et al. (2010). 'CpG Islands Influence Chromatin Structure via the CpG-binding Protein Cfp1'. *Nature* 464.7291, pp. 1082–1086. DOI: 10.1038/nature08924.

Titus, A. J., R. M. Gallimore, L. A. Salas and B. C. Christensen (2017). 'Cell-Type Deconvolution from DNA Methylation: A Review of Recent Applications'. *Human Molecular Genetics* 26.R2, R216–R224. DOI: 10.1093/hmg/ddx275.

Tran, R. K., J. G. Henikoff, D. Zilberman, R. F. Ditt, S. E. Jacobsen and S. Henikoff (2005). 'DNA Methylation Profiling Identifies CG Methylation Clusters in Arabidopsis Genes'. *Current biology: CB* 15.2, pp. 154–159. DOI: 10.1016/j.cub.2005.01.008.

Varma, S. J., E. Calvani, N.-M. Grüning, C. B. Messner, N. Grayson, F. Capuano, M. Mülleder and M. Ralser (2022). 'Global Analysis of Cytosine and Adenine DNA Modifications across the Tree of Life'. *eLife* 11, e81002. DOI: 10.7554/eLife.81002.

Vasimuddin, Md., S. Misra, H. Li and S. Aluru (2019). 'Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems'. *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*. 2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS), pp. 314–324. DOI: 10.1109/IPDPS.2019.00041.

Velasco, G., F. Hubé, J. Rollin, D. Neuillet, C. Philippe, H. Bouzinba-Segard, A. Galvani, E. Viegas-Péquignot and C. Francastel (2010). 'Dnmt3b Recruitment through E2F6 Transcriptional Repressor Mediates Germ-Line Gene Silencing in Murine Somatic Tissues'. *Proceedings of the National Academy of Sciences* 107.20, pp. 9281–9286. DOI: 10.1073/pnas.1000473107.

Vincent, J. J. et al. (2013). 'Stage-Specific Roles for Tet1 and Tet2 in DNA Demethylation in Primordial Germ Cells'. *Cell Stem Cell* 12.4, pp. 470–478. DOI: 10.1016/j.stem.2013.01.016.

Visel, A., E. M. Rubin and L. A. Pennacchio (2009a). 'Genomic Views of Distant-Acting Enhancers'. *Nature* 461.7261, pp. 199–205. DOI: 10.1038/nature08451.

Visel, A. et al. (2009b). 'ChIP-seq Accurately Predicts Tissue-Specific Activity of Enhancers'. *Nature* 457.7231, pp. 854–858. DOI: 10.1038/nature07730.

Walsh, C. P. and T. H. Bestor (1999). 'Cytosine Methylation and Mammalian Development'. *Genes & Development* 13.1, pp. 26–34. DOI: 10.1101/gad.13.1.26.

Wang, X. et al. (2011). 'Residual Embryonic Cells as Precursors of a Barrett's-like Metaplasia'. *Cell* 145.7, pp. 1023–1035. DOI: 10.1016/j.cell.2011.05.026.

Watt, F. and P. L. Molloy (1988). 'Cytosine Methylation Prevents Binding to DNA of a HeLa Cell Transcription Factor Required for Optimal Expression of the Adenovirus Major Late Promoter.' *Genes & Development* 2.9, pp. 1136–1143. DOI: 10.1101/gad.2.9.1136.

Weber, M., I. Hellmann, M. B. Stadler, L. Ramos, S. Pääbo, M. Rebhan and D. Schübeler (2007). 'Distribution, Silencing Potential and Evolutionary Impact of Promoter DNA Methylation in the Human Genome'. *Nature Genetics* 39.4, pp. 457–466. DOI: 10.1038/ng1990.

Wheldon, L. M. et al. (2014). 'Transient Accumulation of 5-Carboxylcytosine Indicates Involvement of Active Demethylation in Lineage Specification of Neural Stem Cells'. *Cell Reports* 7.5, pp. 1353–1361. DOI: 10.1016/j.celrep.2014.05.003.

Whyte, W. A., D. A. Orlando, D. Hnisz, B. J. Abraham, C. Y. Lin, M. H. Kagey, P. B. Rahl, T. I. Lee and R. A. Young (2013). 'Master Transcription Factors and Mediator Establish Super-Enhancers at Key Cell Identity Genes'. *Cell* 153.2, pp. 307–319. DOI: 10.1016/j.cell.2013.03.035.

Wickham, H. et al. (2019). 'Welcome to the Tidyverse'. *Journal of Open Source Software* 4.43, p. 1686. DOI: 10.21105/joss.01686.

Widschwendter, M. et al. (2007). 'Epigenetic Stem Cell Signature in Cancer'. *Nature Genetics* 39.2, pp. 157–158. DOI: 10.1038/ng1941.

Wiencke, J. K. et al. (2017). 'Immunomethylomic Approach to Explore the Blood Neutrophil Lymphocyte Ratio (NLR) in Glioma Survival'. *Clinical Epigenetics* 9, p. 10. DOI: 10.1186/s13148-017-0316-8.

Wilkins, O. M., K. C. Johnson, E. A. Houseman, J. E. King, C. J. Marsit and B. C. Christensen (2019). 'Genome-Wide Characterization of Cytosine-Specific 5-Hydroxymethylation in Normal Breast Tissue'. *Epigenetics* 15.4, pp. 398–418. DOI: 10.1080/15592294.2019.1695332.

Williams, K., J. Christensen, M. T. Pedersen, J. V. Johansen, P. A. C. Cloos, J. Rappsilber and K. Helin (2011). 'TET1 and Hydroxymethylcytosine in Transcription and DNA Methylation Fidelity'. *Nature* 473.7347, pp. 343–348. DOI: 10.1038/nature10066.

Wolffe, A. P. and M. A. Matzke (1999). 'Epigenetics: Regulation through Repression'. *Science (New York, N.Y.)* 286.5439, pp. 481–486. DOI: 10.1126/science.286.5439.481.

Wu, H., A. C. D'Alessio, S. Ito, Z. Wang, K. Cui, K. Zhao, Y. E. Sun and Y. Zhang (2011a). 'Genome-Wide Analysis of 5-Hydroxymethylcytosine Distribution Reveals Its Dual Function in Transcriptional Regulation in Mouse Embryonic Stem Cells'. *Genes & Development* 25.7, pp. 679–684. DOI: 10.1101/gad.2036011.

Wu, H., A. C. D'Alessio, S. Ito, K. Xia, Z. Wang, K. Cui, K. Zhao, Y. Eve Sun and Y. Zhang (2011b). 'Dual Functions of Tet1 in Transcriptional Regulation in Mouse Embryonic Stem Cells'. *Nature* 473.7347 (7347), pp. 389–393. DOI: 10.1038/nature09934.

Wu, H. and Y. Zhang (2015). 'Charting Oxidized Methylcytosines at Base Resolution'. *Nature Structural & Molecular Biology* 22.9 (9), pp. 656–661. DOI: 10.1038/nsmb.3071.

Wu, T. et al. (2021). 'clusterProfiler 4.0: A Universal Enrichment Tool for Interpreting Omics Data'. *The Innovation* 2.3. DOI: 10.1016/j.xinn.2021.100141.

Wu Ct, n. and J. R. Morris (2001). 'Genes, Genetics, and Epigenetics: A Correspondence'. *Science (New York, N.Y.)* 293.5532, pp. 1103–1105. DOI: 10.1126/science.293.5532.1103.

Xu, G. L., T. H. Bestor, D. Bourc'his, C. L. Hsieh, N. Tommerup, M. Bugge, M. Hulten, X. Qu, J. J. Russo and E. Viegas-Péquignot (1999). 'Chromosome Instability and Immunodeficiency Syndrome Caused by Mutations in a DNA Methyltransferase Gene'. *Nature* 402.6758, pp. 187–191. DOI: 10.1038/46052.

Xu, H., J. Chen, J. Cheng, L. Kong, X. Chen, M. Inoue, Y. Liu, S. Kriaucionis, M. Zhao and C.-X. Song (2023). 'Modular Oxidation of Cytosine Modifications and Their Application in Direct and Quantitative Sequencing of 5-Hydroxymethylcytosine'. *Journal of the American Chemical Society* 145.13, pp. 7095–7100. DOI: 10.1021/jacs.3c01663.

Yang, H. et al. (2013). 'Tumor Development Is Associated with Decrease of TET Gene Expression and 5-Methylcytosine Hydroxylation'. *Oncogene* 32.5, pp. 663–669. DOI: 10.1038/onc.2012.67.

Yang, R. et al. (2017). 'The Association between Breast Cancer and S100P Methylation in Peripheral Blood by Multicenter Case-Control Studies'. *Carcinogenesis* 38.3, pp. 312–320. DOI: 10.1093/carcin/bgx004.

Yoder, J. A. and T. H. Bestor (1998). 'A Candidate Mammalian DNA Methyltransferase Related to Pmt1p of Fission Yeast'. *Human Molecular Genetics* 7.2, pp. 279–284. DOI: 10.1093/hmg/7.2.279.

Yu, M. et al. (2012). 'Base-Resolution Analysis of 5-Hydroxymethylcytosine in the Mammalian Genome'. *Cell* 149.6, pp. 1368–1380. DOI: 10.1016/j.cell.2012.04.027.

Zhu, J., F. He, S. Hu and J. Yu (2008). 'On the Nature of Human Housekeeping Genes'. *Trends in genetics: TIG* 24.10, pp. 481–484. DOI: 10.1016/j.tig.2008.08.004.

Zilberman, D., M. Gehring, R. K. Tran, T. Ballinger and S. Henikoff (2007). 'Genome-Wide Analysis of Arabidopsis Thaliana DNA Methylation Uncovers an Interdependence between Methylation and Transcription'. *Nature Genetics* 39.1, pp. 61–69. DOI: 10.1038/ng1929.

Ziller, M. J. et al. (2013). 'Charting a Dynamic DNA Methylation Landscape of the Human Genome'. *Nature* 500.7463 (7463), pp. 477–481. DOI: 10.1038/nature12433.