Searching for Cancer Biomarkers in the Human Epigenetic Landscape



Felix Jackson

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Hilary 2024

Declaration

I declare that this thesis was composed by myself and that the work contained is my own except where explicitly stated in the text. This work has not been submitted for any degree of professional qualification except as specified.

Felix Jackson, Hilary term 2024

Acknowledgements

About half way through this DPhil, it became clear to me that this research could go on forever. I think the people around me are all glad that it hasn't. I am extremely grateful to everyone who has supported me over the last five years, and am excited to spend time with each of them without fretting about being away from my laptop.

First and foremost, I am very grateful to my supervisors Professor Chunxiao Song and Professor Thomas Lukasiewicz. Without their generous support and guidance, these research projects would not have been possible. I have learnt a great deal from four years of lab meetings with Chunxiao and his lab, and he continues to be an inspiring research mentor and role model for me. Secondly, I'd like to thank Dr Jingfei Cheng, who has tolerated me distracting her from her research with endless questions about data pipelines, and has somehow kept her chill despite me asking the same questions about bash hundreds of times. Through our collaborations she has taught me a great deal. Thanks also to Masato for being a fantastic experimental collaborator. Clinical research is a team sport, and is not possible without many dedicated people holding up the research infrastructure behind the scenes. I am grateful to the BMRC team who have been consistently helpful, and keep the medical research cluster afloat despite our best attempts. I am also indebted to all the patients who agreed to provide data for this research, without whom none of this would be possible.

Outside of academia, I am lucky to have been surrounded by a brilliant group of friends who have kept me ~~distracted~~ entertained with many fun adventures in Oxford and further afield. I particularly appreciate my fantastic bandmates who have accompanied me for a musical journey running parallel to this research journey, and my surfer friends who have kept the ocean close.

I owe most of all of this to my amazing parents, who are a constant source of wisdom and support, as well as the occasional friendly nudge. I'm grateful to my Granny, the family matriarch, and my excellent brothers who have always led the way, keeping me busy with outdoor pursuits and a growing number of nieces and nephews. Finally, I'm grateful to my partner Saran, who has shared the front seat with me on this journey. Hopefully soon we can go to the pub.

Abstract

Methylation and hydroxymethylation are key epigenetic regulators of the genome. Without changing the underlying sequence, these epigenetic modifications allow cells to orchestrate highly specialised gene expression programs, up-regulating genes with important tissue-specific functions and silencing genes used elsewhere. Methylation patterns across the genome are highly tissue-specific, which makes them useful candidate biomarkers for a variety of diagnostic tests. Methylation changes are also emerging as a consistent early harbinger of malignancy in cancer. There is a pressing need for non-invasive cancer diagnostic tests, and therefore intense interest in the development of methylation-based cancer diagnostics.

In this thesis, I present novel methods and data focused on gaining a better understanding of genome-wide methylation and hydroxymethylation patterns, and explore how these insights can be utilised for non-invasive cancer diagnosis. I collate a comprehensive set of genome-wide tissue-informative markers from public data, then use these markers to perform deconvolution of cell-free DNA, obtained from liquid biopsy of cancer patients. I identify significantly increased liver tumour contribution in liver cancer patients, and show the inferred tissue contribution can be used to distinguish cancer patients from non-cancer controls. In clinical contexts, patient data is generally scarce and a limiting factor for diagnostic models. I devise new machine-learning based cfDNA deconvolution and classification algorithms with more efficient data use and improved diagnostic accuracy.

The availability of high quality training data is a major bottleneck for these models, and in the second half of this thesis I focus on collating and analysing a new resource dataset: genome-wide methylation and hydroxymethylation signatures for 22 healthy tissues and 10 tumour types at the resolution of individual CpGs. I develop a novel spatial segmentation method to disaggregate the genome into functional building blocks of methylation and hydroxymethylation, then identify tissue and tumour informative regions within these blocks. Finally, I explore the downstream functional consequences of methylation and hydroxymethylation by building a gene expression prediction model that can accurately predict gene expression profiles in unseen healthy and tumour tissues, using these epigenetic marks alone.

Contents

1 Introduction **1**

1.1 The biology of DNA methylation 1
1.2 Hydroxymethylation as a distinct epigenetic marker 5
1.3 The case for early cancer detection through non-invasive diagnostics 5
1.4 Epigenetic analysis of cell-free DNA and its utility in cancer diagnostics 6
1.5 Deconvolution of cell-free DNA into its tissues of origin 10
1.6 Methylation sequencing methods 11
1.7 Challenges of analysing biological sequence data 13
1.8 Availability of public epigenetic data 15
1.9 Machine learning methods for sequence data 15
1.10 List of publications 18

2 Background **19**

2.1 The nature of methylation sequencing data 19
2.2 Identifying differentially methylated regions 20
2.3 Modelling assumptions for methylation data 23
2.4 Context is key: defining the background in DMR-finding algorithms 25
2.5 Defining DNA methylation signatures 26
2.6 Measuring tissue specificity of candidate DMRs 26

3 Tissue deconvolution of cell-free DNA methylation data **29**

3.1 Abstract 29
3.2 Introduction 30
3.3 Assembling a genome-wide tissue methylation reference atlas 36

3.3.1 Unsupervised clustering of tissue samples . . . 38
3.3.2 Identification of tissue-specific differentially methylated regions . . . 39
3.3.3 Evaluating DMR tissue-specificity . . . 42
3.4 Tissue deconvolution of cell-free DNA using a methylation reference atlas . . . 46
3.4.1 Cell-free DNA patient cohort . . . 46
3.4.2 Tissue deconvolution by non-negative least squares regression 47
3.4.3 Tissue deconvolution of cfDNA using CelFiE . . . 49
3.5 Developing a deep learning autoencoder for cfDNA tissue deconvolution . . . 53
3.5.1 Training on simulated data . . . 54
3.5.2 Deconvolution of patient cell-free DNA . . . 55
3.5.3 Reconstruction of cell-type specific methylation signatures . 57
3.6 Discussion . . . 59

4 Multimodal classifiers of cell-free DNA 64
4.1 Abstract . . . 64
4.2 Introduction . . . 65
4.3 Classifying cell-free DNA with a multimodal classifier . . . 67
4.4 Developing a read classifier of cell-free DNA . . . 69
4.5 Evaluating cfDNA read classifier on patient data . . . 72
4.6 Discussion . . . 75

5 Constructing a reference atlas of tissue methylome and hydroxymethylome signatures 79
5.1 Abstract . . . 79
5.2 Introduction . . . 80
5.3 Constructing a methylation and hydroxymethylation reference atlas 83
5.3.1 Tissue sample quality control and clustering . . . 84
5.4 Segmenting methylation and hydroxymethylation into functional building blocks . . . 90
5.4.1 Evaluating existing segmentation methods . . . 90

5.4.2 Benchmarking methods on synthetic data 90
5.4.3 Testing segmentation algorithms on tissue methylation data 93
5.4.4 Segmenting mC and hmC across all tissue samples 97
5.4.5 Segmenting the cancer methylome and hydroxymethylome . 98
5.4.6 Unsupervised clustering of tissue samples over blocks 98
5.5 Identifying tissue-specific differentially methylated and hydroxymethylated blocks 100
5.5.1 Identifying tissue-specific differentially methylated blocks . . 101
5.5.2 Downstream interpretation of tissue-specific DMBs 105
5.5.3 Identifying cancer-specific differentially methylated blocks . 108
5.5.4 Deconvolution of cell-free DNA using tissue informative markers 111
5.5.5 Searching for pan-cancer DMBs 112
5.5.6 Automated feature selection methods for tissue-specific DMRs 114
5.5.7 Non-negative matrix factorisation to identify tissue-specific DMBs 116
5.6 Discussion 119

6 Predicting gene expression using genome-wide patterns of cytosine methylation and hydroxymethylation 123
6.1 Abstract 123
6.2 Introduction 124
6.3 Obtaining Linked Epigenetic and Gene Expression Data 126
6.4 Correlation of methylation and hydroxymethylation with RNA-seq transcript abundance 129
6.5 Predictive models of gene expression 131
6.5.1 Model architecture and training 133
6.6 Feature importance of gene expression prediction model 137
6.7 Validating predictive model on new cancer gene expression data . . 139
6.8 Discussion 141

7 Discussion **144**

7.0.1 Conclusions and key contributions 152

7.0.2 Future work 153

A Appendix **156**

A.0.1 Cell-free DNA patient cohort 156

A.0.2 TAPS tissue atlas QC and clustering 160

Bibliography **168**

List of Figures

2.1 An example of finding DMRs on synthetic data 26

3.1 Schematic illustration of the cfDNA tissue deconvolution problem. Each tissue sheds DNA into the blood containing a unique methylation signature, which then circulates as cell-free DNA. The resulting cfDNA can be extracted from blood and sequenced non-invasively, but the observed mixture must be separated into its constituent components. 31

3.2 Formalisation of the cfDNA tissue deconvolution problem in matrix form. k tissues shed cfDNA into the blood at tissue-specific rates captured in matrix F, forming a cfDNA pool which is the observed mixture D. Deconvolution methods separate this observed mixture into a matrix of proportions A and a reference T. 32

3.3 The number of samples for each tissue in our final reference dataset after filtering samples, split by source dataset. Liver and liver tumour samples marked 'Other' were obtained from personal correspondence with Li *et al.* 38

3.4 All tissue samples in the final reference atlas projected into 2D space using a t-SNE embedding of methylation across all Enhancers. . . . 39

3.5 The key DMR filtering criteria I used, illustrated at a single genomic locus containing a (real) liver DMR. Arrows show the median difference and the variance of the target group (liver in this case) and background group. 41

3.6 A heatmap of the final set of DMRs obtained from the reference atlas, after optimising the DMR calling algorithm to extract maximally tissue informative regions. 44
3.7 A t-SNE projection of the tissue atlas intersected on the final set of enhancer DMRs, which are used for downstream tissue deconvolution. 45
3.8 Metadata for this cfDNA patient cohort. Left: age and sex distribution in each of the five groups. Middle: cancer stage distribution for the HCC and PDAC cancer patients. Right: mean genome modification for the two cancer types compared to non-cancer controls. 47
3.9 Left: mean tissue contribution in control group, dominated by blood and immune cell types. Middle: liver tumour contribution across the three groups, demonstrating a statistically significant increase in liver tumour contribution in HCC. Right: T cell contribution across three groups, demonstrating a statistically significant increase in T cell contribution assigned to PDAC cfDNA but not to HCC cfDNA. 48
3.10 Mean tissue contribution across the three groups as inferred by NNLS 49
3.11 Mean tissue contribution across the three groups as inferred by NNLS 51
3.12 Mean tissue contribution across the three groups as inferred by Celfie, with one unknown stipulated. 51
3.13 Boxplot comparing tissue contribution across healthy samples, as estimated by NNLS and CelFiE algorithms. Boxplots show median and inter-quartile range, with outliers labelled as dots. 52
3.14 Boxplot comparing tissue contribution across HCC samples, as estimated by NNLS and CelFiE algorithms. Difference between model estimates for most tissues were statistically significant, except for osteoclasts and monocyte. 52
3.15 Autoencoder schematic, showing input, output and model loss. Also shown are the two separate losses as used in model training. 54
3.16 Cell type contribution averaged across each of the two groups, liver cancer (HCC) and non-cancer controls. 56

3.17 Left: Differences in inferred T cell and hepatocyte contribution between HCC and non-cancer cfDNA samples. Right: ROC curve for random forest and gradient boosting classifiers trained to distinguish HCC from non-cancer samples using the inferred contributions. Random forest AUC: 0.94 with std. dev 0.11; gradient boosting AUC: 0.81 with std dev 0.09 57
3.18 Pearson correlation of reconstructed cell-type specific methylation signatures with true reference signature. 58

4.1 A schematic of cell-free DNA fragments found in peripheral blood, aligned to a gene of interest. These fragments contain three distinct modalities of information: DNA sequence, methylation pattern (shown here on CpGs) and their variable fragment length. The genomic locus that each fragment originates from is also informative. 65
4.2 (A) Heatmap showing individual model performance on multicancer prediction and the predicted probabilities for each patient. Each vertical column is a patient. Detection yes/no refers to whether patients are correctly classified or misclassified on the basis of a particular feature. Predicted score means the probability of classifying the patients to a specific group based on a particular feature. (B) Schematic detailing the method of integrating multiple features (DNA methylation, tissue contribution, and cfDNA fragmentation) extracted from cell-free TAPS data for multicancer prediction. (C) Overall accuracy and confusion matrix showing the actual and predicted patient status for the ensemble model 68
4.3 Schematic for assigning model scores to individual fragments, which are then aggregated into a patient level classification. 69
4.4 The distribution of cfDNA fragment scores across healthy controls (green) and liver cancer patients (red) in the test set. Each line represents the distribution over all scores for a single patient's cfDNA sample, where the score for a single fragment is the probability that fragment originated from cancer. This data is patient groups 1 and 2. 73

4.5 Left: patient risk scores for HCC and healthy cfDNA CancerDetector, taken by integrating all assigned model probabilities from individual fragments. Right: ROC curve generated for distinguishing TAPS cfDNA samples from HCC and non-cancer controls from patient-level risk scores, AUC = 0.75 74
4.6 The distribution of cfDNA fragment risk scores stratified by fragment length, for a single liver cancer patient (downsampled for legibility). Methylation state (normalised by fragment length) is represented by marker color, where a lighter color means higher methylation. 75

5.1 Schematic of all tissue samples collected for our reference atlas . . . 84
5.2 Mean coverage across all CpGs for all samples in the two experiments, TAPSbeta and CAPS. 85
5.3 Mean methylation and hydroxymethylation ratios for all samples . . 86
5.4 t-SNE projection of final reference atlas after removing outliers . . . 87
5.5 Mean global levels of mC (top) and hmC (middle) across all CpGs in our reference. Bottom panel shows overall modification as calculated using MLML. 89
5.6 An example of the simulated methylated data used for benchmarking algorithms. The black lines indicates the true positions of the blocks in the simulated data, and the dashed green lines indicate the assigned boundaries found by each of the four segmentation algorithms. 91
5.7 Comparing performance of the various segmentation algorithms. Left: F1 score of each segmentation algorithm, comparing the number of true simulated blocks correctly identified against the number of false positives across different simulated scenarios. Right: the time taken for each algorithm to run on different data scales. 94
5.8 HMM segmentation of mC (top) and hmC (bottom) around the SERBP1 and TNN13 genes. 95

5.9 Comparison of different HMM parameter values over a single region surrounding the A1BG gene. The upper two panels show the resulting segmentation for two distinct values for HMM hidden states. The lower panel shows the BIC and silhouette scores for these parameter values, and the composite score from combining the two with equal weighting. 96
5.10 Median block length (left), and median number of CpGs per block (right) for mC and hmC segmentation (top and bottom respectively) 97
5.11 A comparison of segmenting the same region using the same algorithm, but on tumour methylation data (right) compared to healthy samples (left). 98
5.12 Dendrogram clustering of tissue samples using methylation and hydroxymethylation blocks. NMI and ARI scores indicate the quality of the clustering. 99
5.13 Illustration of new DMB finding approach for a real Pancreas DMB. The new appproach ranks regions by the methylation delta between the 75th percentile of the target group and the 2.5th percentile of the background group, for hypomethylated DMBs. My previous approach used the methylation delta between group medians. . . . 102
5.14 Top 300 hypermethylated (left) and hypomethylated (right) mC DMBs for each healthy tissue in our tissue atlas 103
5.15 Top 300 hypermethylated (left) and hypomethylated (right) hmC DMBs for each healthy tissue in our tissue atlas 103
5.16 All TAPSbeta DMBs for healthy tissues, split into hyper DMBs (top) and hypo DMBs (bottom), and coloured by whether or not each DMB is assigned a Bonferroni corrected significant p-value. . . 105

5.17 Metagene plot calculated across all genes and all tissues, showing the distribution of DMBs across the gene body. The y-axis shows the number of DMBs linked to tissue-specific genes, normalised by the number of tissue-specific genes. This is shown separately for hyper-hydroxymethylated DMBs (left) and hypomethylated DMBs (right). The x-axis indicates the transcription start site (TSS) and transcription end site (TES), plus 10-kilobases in either direction. The region between TSS and TES is considered the gene body. . . . 106
5.18 Enrichment of top 300 DMB-linked genes in tissue-specific genes downloaded from EnrichR, with EnrichR tissues on y-axis and atlas tissues on x-axis. The same method is shown for mC hypo DMBs (left) and hmC hyper DMBs (right). For tissue-specific genes from solid tissues, the original dataset was GTEx Tissues V8 2023; for immune cells the original dataset was PanglaoDB Augmented 2021. 107
5.19 Enrichment between hypo mC DMBs and H3K27Ac (left), and hyper hmC DMBs (right) for a variety of tissues. The profile of histone markers across many other tissues are shown for comparison. Histone enrichment for the target tissue is shown with a solid line, and other non-target tissues are shown with dotted lines. For example, this indicates that around heart-specific DMBs from our tissue atlas, we observe enrichment in H3K27Ac histone marks for heart tissue, but not for other tissues. 108
5.20 Hyper and hypomethylated tumour specific DMBs for methylation . 109
5.21 Hyper and hypomethylated tumour specific DMBs for hydroxymethylation 109
5.22 DMBs for each tumour type ranked by value of the methylation delta between assigned quantiles, coloured by p-value significance in Bonferroni corrected Welch's t-test. The red text shows the total number of DMBs. 110
5.23 Enrichment of tumour DMB-related genes within up-regulated genes in tumour samples. Left shows enrichment for hypo mC DMBs and right shows the same for hyper hmC DMBs. 111

5.24 Predicted tissue contribution in HCC and non-cancer controls, using deep neural network deconvolution method re-trained on simulated cfDNA mixtures from this new reference atlas. 112
5.25 Proportion of shared DMBs for each tumour type 113
5.26 Proportion of shared DMBs for each tumour type 114
5.27 Methylation ratios of all samples in a hyper DMB and a hypo DMB shared between 8 tumour types. Tissue types classed as containing this DMR are starred. 114
5.28 Left: Heatmap showing methylation values for pan-cancer features identified by RF classifier. Right: PCA demonstrating separation of samples on these genomic features. 116
5.29 Correlation of each LMC with healthy tissue samples used to perform NMF 117
5.30 Methylation ratio for the top 50 highest weighted genomic features for each LMC in the above NMF. 118

6.1 Probability distribution of methylation and hydroxymethylation across promoters, enhancers and gene bodies of the 15,894 covered genes in this dataset, calculated using kernel density estimation. For clarity, four representative tissues/cell types are highlighted to show the distinct distributions of solid tissues, brain, tumour types, and blood cell types. Remaining tissues are shown in grey. 128
6.2 Pearson correlation of methylation and hydroxymethylation across promoter, enhancer and gene body genomic elements of all 15,894 genes. This figure illustrates correlation for a single tissue (lung), which is a representative example of the general trend in correlation across all tissues in the tissue atlas. A linear line of best fit and Pearson correlation value is also shown. 130

6.3 Pearson correlation of methylation and hydroxymethylation across promoter, enhancer and gene body genomic elements of all, separated by tissues. The left-hand plot shows this correlation for all 15,894 genes in our filtered dataset, then moving right we see the same correlations for the subset of housekeeping and tissue specific genes. Beneath each plot is the mean correlation value for each feature across all tissues. On the right hand side, the cross-correlation of gene body hydroxymethylation of every tissue in the tissue atlas with every RNA-seq tissue expression profile, thus demonstrating the tissue specificity of this epigenetic mark. 131
6.4 A single example of an input data array for a random gene and tissue, to illustrate the input data for our ConvNet model. 132
6.5 Methylation, hydroxymethylation and unmethylated fractions surrounding the TSS of a randomly selected gene (SKI), binned into 1kb bins for 20kb upstream and downstream of gene body. The gene body is divided into 20 equal length segments. This is used as input data to a predictive model, where the target is RNA-seq Gene expression (logTPM) as shown on the right. 135
6.6 Model performance of gene expression prediction model, evaluated on all genes for a single held-out tissue. R-squared plots for best and worst performing models are shown on the top panel. The lower panel shows the distribution of R-squared values. 136
6.7 SHAPley feature importance for mC, hmC and umC across all tissues for the 1kb model (left) and the 2kb model (right). 138

6.8 Z-score of Pearson correlation for model predictions with two independent cancer gene expression datasets. GBM was missing for the second dataset. For each tumour type in our reference, I identify the best matching cancer type with public gene expression data, for example brain tumour and glioblastoma (GBM). I then predict gene expression for all genes using mC and hmC distribution in brain-tumour and correlate these predictions to observed GBM gene expression. To benchmark this score and investigate its tumour specificity, I also compare gene expression model predictions to all other cancer types. 140

A.1 Number of remaining CpGs after filtering all rows of the reference atlas by different coverage thresholds. 165

A.2 Copy number variation plot for two heart samples generated by ichorCNA. The upper panel is representative of a 'normal' CNA plot, whereas the lower plot shows wide heterogeneity indicating sample quality issues. 165

A.3 Count of CpG sites in each sample with coverage less than 4. Outlier samples with many low coverage sites are labelled in red. 166

A.4 Deconvolution of our reference atlas on a recently published methylation reference atlas. There is not a one-to-one correspondence between tissue types in the two references, but they have been matched as closely as possible. 167

Chapter 1

Introduction

1.1 The biology of DNA methylation

The genomic DNA sequence in all cell types of the human body is identical, yet our cells differentiate into a wide variety of stable phenotypes, and perform distinct, highly specialised tasks in each organ system. Epigenetic modifications to DNA enable this. Cytosine methylation is one such epigenetic modification that occurs pervasively at CpGs across the genome, which orchestrates gene expression by regulating transcriptional access to chromatin. DNA methylation is a means of annotating the genome, a molecular bookmark on genetic text, which directs the cell towards parts of the genome that should be transcribed and expressed, and keeps other parts of DNA folded up and suppressed. Adding methyl groups to cytosines in DNA changes the local chemical environment, causing chromatin to fold up and resist transcription by physically blocking access to transcribing proteins at that locus. Conversely, demethylation increases chromatin accessibility to the transcription machinery, thus encouraging gene expression at that genomic locus. Genes that are preferentially demethylated to cause expression in kidney cell types will be methylated and suppressed in other tissues such as heart. Hepatocyte specific genes are generally methylated in all cell types outside the liver. In this way, the methylome of each tissue sets up a tissue-specific gene expression program leading to stable cellular phenotypes, adapted to their role [1, 2]. A cell's methylation state - the bookmarks in the text of the genome - is somatically heritable, yet unlike DNA sequence can adapt in response to a changing external

environment [3].

Human genetic information is stored in the nucleus as chromatin, which is DNA wrapped around histone proteins. The 3D structure of chromatin directly affects gene expression by controlling the accessibility of DNA to proteins involved in gene transcription. Epigenetic modifications influence gene expression by binding to DNA or histones and altering the accessibility of DNA to gene expression machinery. Unlike the genetic code, these epigenetic marks vary widely across cell types and fluctuate through time. Taken together these epigenetic marks constitute a second layer of information in each cell, sitting atop the genome and acting as a type of molecular 'bookmark', directing gene transcription machinery to certain genomic loci that lead to distinct cell identities. DNA methylation - the addition of a methyl group to the C5 position of cytosine by various DNA methyltransferase (Dnmt) enzymes - is a well established example of this phenomenon.

Methylation of DNA was discovered alongside Avery's identification of DNA as the genetic material in the 1940s [4], but its role in regulating gene expression was not discovered until 1975 [5]. Some molecular details are still missing, but cytosine methylation is known to regulate gene transcription by at least two distinct mechanisms: first by binding DNA and physically obstructing DNA-binding transcriptional activator proteins, most commonly at gene promoters [6, 7]; second by acting as a signalling molecule and recruiting methyl-binding proteins, which subsequently recruit repressive protein complexes [8, 9].

The majority of cytosine methylation occurs at sites of cytosine-guanine dinucleotides, known as CpG sites; although non-CpG methylation does occur, I focus on CpG methylation throughout this thesis. These are generally depleted across the genome due to their mutagenicity, but are enriched in certain clusters known as CpG islands, whose evolutionary conservation implies their functional importance [10]. CpG islands (CGIs) are present in around 70% of gene promoters [11], the sites of gene transcription initiation. Whilst the majority (60-80%) of CpG sites in the genome are methylated, CpG islands in gene promoters are generally demethylated and exhibit a dynamic range of methylation states linked to accessible chromatin and gene transcription [12, 13, 9]. The edges of these elements are known as CGI shores, and exhibit highly conserved tissue-specific methylation patterns where increased methylation is correlated with reduced gene expression

[14]. Overall, although molecular details are still emerging, CGI methylation at gene promoters has been consistently shown to repress gene expression by blocking transcription factor binding and recruiting repressive CpG-methyl binding protein complexes.

CpG methylation serves several important functions in the development and maintenance of a stable genome in mammals, most notably X chromosome inactivation [15], genomic imprinting [16, 17], and the silencing of transposable elements. These are repetitive elements that make up more than half of the genome, and are termed "jumping genes" due to their ability to insert themselves at new genomic loci, where they disrupt normal gene transcription. The insertional mutations caused by these elements have been linked to diseases such as Duchenne muscular dystrophy, hemophilia and cancer [18], so hypermethylation of these elements is essential for long-term genome stability [19]. CpG methylation is symmetric across both DNA strands, and is propagated through cell divisions [13]

The best characterised role of CpG methylation is in repressing gene transcription, and an abundance of evidence links CpG methylation to a repressive chromatin state and gene silencing. However as genome-wide mapping of methylation has become possible, the emerging picture is that the effects of cytosine methylation are highly dependent upon the genomic context. For example, there is some evidence that methylation of gene bodies is positively correlated with gene expression, and may act as a transcriptional activator [20]. There is no consensus mechanism, but a popular theory is that gene body methylation suppresses transcription at intragenic cryptic promoters, which would otherwise serve to interfere with expression of that gene [21], and may also affect alternative splicing [12].

Enhancers are cis-acting DNA regulatory elements that influence gene transcription by recruiting transcription factors from a distance, between 100 base-pairs (bp) to as much as two megabases away from transcription start sites (TSS) [13, 22]. The locations of enhancers in the human genome are generally defined by the presence of two distinct chromatin histone marks: H3K4me1 and H3K27ac [23]. Enhancers are emerging as an important regulator of tissue specific gene expression, and exhibit a wide range of methylation states which seem to correlate with tissue-specific expression. Experiments present mixed reports on how methylation state of enhancers relates to gene expression, but the overall picture is that

active tissue-specific enhancers are hypomethylated in that tissue type, to a low but non-zero methylation level [24, 25, 26, 27].

Just as the methylation landscape of a cell regulates cell activity by modulating gene expression, methylation changes play an important role in the development of cancer. Cancer is fundamentally a disease of dysregulated gene activity, where overactive oncogenes and underactive tumour suppressors tip the delicate balance of cellular homeostasis towards sustained proliferation and reduced cell death. Somatic DNA mutations are still considered the primary mechanism here, but increasing evidence supports the importance of epigenetic changes to this cellular reprogramming [28]; "nonmutational epigenetic reprogramming" was named as an emerging hallmark of cancer in the 2022 update of Hanahan & Weinberg's seminal Hallmarks of Cancer series [29]. Methylation changes were first detected in cancer in the early 1980's, with widespread DNA hypomethylation in cancer cell lines, hypomethylation at some oncogenes in primary tumours and hypermethylation at CpG islands [30, 31, 1]. Many more datapoints have since emerged across numerous cancer types, indicating at least four distinct mechanisms. First, aberrant hypermethylation of CpG Islands in promoters of tumour suppressor genes, such as hypermethylation of the BRCA1 gene in breast and ovarian cancer [32], VHL in renal carcinoma [33], and STK11 in colon cancer [34]. Second, the aberrant hypomethylation of genomic elements which are usually suppressed by methylation, such as SINE and LINE1 retrotransposons in prostate and skin cancer [35, 36]. Thirdly, hypomethylation occurs at promoters of oncogenes, whose expression is tightly regulated in healthy cells: examples include hypomethylation of ELMO3 oncogene in lung cancer and IRX1 in osteosarcoma [37, 38]. Finally enhancer elements, although not yet fully mapped out, are emerging as important regulators of cancer specific gene expression. Enhancer methylation can block access of transcription factor machinery to DNA, effectively decomissioning them, whilst aberrant hypomethylation in enhancers can expose the underlying transcription factor binding motifs to activate oncogene expression [39].

1.2 Hydroxymethylation as a distinct epigenetic marker

In 2009 it was discovered that the same C5 position of cytosine nucleotides can also be hydroxymethylated, catalysed by TET enzyme assisted oxidation of methylcytosine as part of the DNA demethylation pathway [40]. Hydroxymethylated cytosine (hmC) is a known intermediate in the DNA demethylation pathway, but there is emerging evidence that hmC may also act as a stable and functionally distinct epigenetic marker [41, 42, 43]. Bisulfite sequencing, the most widely used methylation sequencing method, cannot distinguish between mC and hmC, so the distinct distribution and role of hmC has hitherto been poorly characterised. Genome wide mapping of 5hmC shows it is enriched in gene bodies and tissue-specific enhancers, and its presence is generally associated with increased transcriptional activation [44, 41, 45, 46]. Relative to methylated cytosine, hydroxymethylated cytosine levels are low in most somatic tissues: generally fewer than 10% of CpGs are hydroxymethylated, and fewer again in blood and immune cell types [47, 44, 43]. A notable exception is in brain and central nervous system tissues, where up to 25-40% of CpGs are hydroxymethylated [48, 49], which has interesting implications on the role of methylation and hydroxymethylation in epigenetic encoding of memory. A global loss of hydroxymethylation is generally observed in cancer [50], although the exact role of this distinct epigenetic marker in cancer aetiology has yet to be established.

1.3 The case for early cancer detection through non-invasive diagnostics

Early detection is the most effective and cost-efficient way to cure cancer. Average 5-year cancer survival at early stage is 91%, while average 5-year survival at late stage is 26% [51]. Unfortunately around 50% of cancers are only detected at an advanced stage [52]. Early detection can only be achieved by screening with highly sensitive and specific diagnostic tests, which has led to the development of many companies and research groups all chasing the same end-goal. Despite the consequent explosion of new experimental and computational methods, no

method has yet proved consistently successful at detecting cancer in its early, highly-treatable stages [53, 54, 55].

Any screening test must be sufficiently non-invasive to be applied at a population-level without causing more harm than good. This simple fact fundamentally limits early cancer detection, as data from the majority of bodily tissues cannot be accessed in a non-invasive manner. Exceptions to this rule include tissues such as lungs, brain and skin, which can be accurately imaged using various frequencies of electromagnetic radiation, thus facilitating effective population-wide screening: the National Lung Cancer Screening Trial (NLST) in the United States demonstrated a 20% reduction in lung cancer deaths in the low-dose CT arm of the trial (n=15655) [56]. This result validated the potential benefits of implementing screening for high-risk populations. There is currently limited rollout of screening in the US and UK for five cancer types: breast, cervical, colorectal, lung, and prostate [54]. There is an unmet need to expand screening to more cancer types, and non-invasive diagnostic tests can help meet this need.

For the majority of remaining tissues which cannot be interrogated by imaging, other non-invasive diagnostic methods have been explored, in the hope of replacing surgical biopsies as the principal source of data. This has led to the development of a molecular diagnostics field, which aims to detect subcellular macromolecules such as proteins, RNA and DNA in bodily fluids. Of particular interest in recent years is the detection of cell-free DNA (cfDNA) by 'liquid biopsy' of blood plasma [57, 58].

1.4 Epigenetic analysis of cell-free DNA and its utility in cancer diagnostics

DNA fragments are shed into the peripheral blood supply as a result of cell death. These fragments are released following nucleosomal digestion during apoptosis, and exhibit a median length of 167bp: the length of one full DNA loop around a nucleosome [59, 60]. DNA fragment shedding occurs in the majority of somatic tissues, thus allowing tissue-specific information to be gleaned by analysing the genetic and epigenetic patterns in these circulating DNA fragments. Cell-free DNA fragments can be detected in blood plasma and analysed following simple blood

draws, which represents a promising avenue towards non-invasive diagnostics. This approach has already been clinically validated for non-invasive neonatal diagnostics and transplant monitoring [61].

The relative rate of tissue-specific DNA shedding and subsequent clearance is not known. The rate of DNA shedding into the blood fluctuates widely according to cell cycle events, and cfDNA has a half-life of approximately two hours in the blood [62]: given the highly dynamic nature of this biomarker, it is likely that fragmentation patterns reflect metabolic changes in the body. The mechanisms of cfDNA release and clearance are not well established, but a working hypothesis is that a background rate of DNA shedding occurs from each tissue in healthy patients, and tissue-specific physiological changes such as inflammation, tissue injury or vigorous exercise result in increased shedding [2, 63]. Cell-free DNA has been used as a marker of inflammation, to measure the body's response to infections and sepsis, and the recent SARS-CoV2 pandemic [64, 65]. The temporal fluctuation of cfDNA is often overlooked, but this should be considered a dynamic system with continuous release and subsequent clearance, where the relative rates of tissue shedding determine the observed mixture [62]. The majority of cell-free DNA molecules originate from plasma adjacent cell types such as white blood cells, endothelial cells, and blood adjacent tissues such as liver hepatocytes [65]. However, the relative proportions of cfDNA from different tissues reflect tissue health and cell turnover, which makes it a diagnostically useful biomarker.

Circulating tumour DNA (ctDNA) can also be detected in blood plasma: current evidence suggests that cancerous tissue sheds DNA fragments into the blood at a higher rate than its surrounding healthy tissue. Cancer tissue has been widely found to release cell-free DNA at a higher rate, and elevated cfDNA concentrations can be detected in blood of late-stage cancer patients [66, 67, 59]. A cfDNA-based prognostic tool would offer numerous advantages over current oncology practises which rely on biopsy for prognostic biomarkers: its non-invasive nature would reduce the cost, inconvenience and risk of diagnostic tests, thus increasing early detection and allowing for serial measurements. This in turn facilitates detection of tumour burden, treatment response and evolved resistance. Furthermore, tumour DNA is expelled into the blood from diverse regions within the tumour

microenvironment, so ctDNA analysis offers the chance to capture a greater degree of tumour heterogeneity and reduce the sampling bias present in biopsies [63].

Despite considerable interest however, few ctDNA-based diagnostic tools have demonstrated sufficient accuracy to be adopted into clinical use. Over 15,000 articles have been published on cfDNA biomarkers, but only 14 have been adopted for clinical use and only two have been FDA-approved [68]. The development of high-depth next-generation sequencing before base resolution epigenetic sequencing was possible meant initial research focused on mutation-based cfDNA biomarkers. However, a key limiting factor was low sample recovery. The low quantities of sample DNA combined with large intra- and inter-tumour variability in somatic mutation profiles limited the reliability of mutations as a diagnostic biomarker in this context. Cancer-specific mutations have low copy number in the genome, especially at an early stage, and circulating tumour DNA makes up a small fraction of total cfDNA, among contributions from blood cells and other tissue types: a 10ml blood sample is thought to contain less than one genome equivalent of ctDNA. Identifying cancer-specific mutations in this context is often compared to a "needle in a haystack" scenario.

By contrast, DNA methylation is a hallmark of cancer that has become increasingly prominent in mechanistic models of cancer aetiology, due to its frequent implication early in disease progression [69, 28]. For example, aberrant promoter methylation is one of the earliest and most consistently detectable signs of neoplastic growth, which has been used to define leukaemia subtypes [70]. Relying on DNA sequence of cfDNA alone, it is difficult to identify the tissue or tumour origin of specific aberrations. By contrast, cfDNA methylation carries tissue-specific signatures, so changes can be mapped back to their tissue of origin. These patterns can be detected on cfDNA in blood, meaning methylation patterns can be used to distinguish tumours of distinct origin, opening the door to use of liquid biopsy for multi-cancer tests [71]. Cancer is a disease of profound physiological and metabolic changes, and by using cfDNA methylation to map the tissue origins of cfDNA, we can identify these broader physiological changes outside the tumour itself, such as increased inflammation and tissue turnover.

There is considerable industry interest in the use of cell-free DNA liquid biopsy for early cancer detection, and several prospective clinical trials are currently un-

derway. These trials focus on three distinct properties of cfDNA: cancer-specific mutations, methylation signatures, and cfDNA fragmentation profiles. The majority of these rely on identifying mutations in ctDNA, such as CancerSeek and Guardant360 from GuardantHealth [54]. These tests rely on targeted deep sequencing on pre-defined mutation marker panels, due to the combination of low allele frequencies of tumour-specific mutations in cfDNA and the prohibitively high cost of deep genome sequencing. The CancerSeek test demonstrated most promising results in a pilot study on 1005 individuals, but a subsequent prospective trial demonstrated a sensitivity of 27% across all cancer types (specificity 98.9%). Two promising trials rely on cfDNA methylation: PATHFINDER by Grail and a cfMeDIP-seq based approach [54, 53]. The Grail cfDNA methylation test ("Galleri") is currently the largest scale and most promising pan-cancer diagnostic test in this space. However, even this method has so far demonstrated low sensitivity for detection of early stage cancers, and demonstrated an overall sensitivity of 29% at a specificity of 99.1% [72].

Cell-free DNA fragmentation profiles are non-uniform across the genome, and reflect the nucleosome positioning and chromatin accessibility in the tissue of origin [66]. By inferring chromatin accessibility we can estimate the gene expression program in the tissue of origin, and thus map cfDNA back to its contributing tissues [73, 74, 75]. The fragment length distribution of cfDNA also changes noticeably in cancer, with a shorter mean length of cfDNA fragments and a wider dispersion of lengths. Genome-wide sequencing combined with fragmentation profiling has therefore been used to successfully distinguish cancer patients from healthy controls [59]. A number of further prospective trials are therefore being carried out based on cfDNA fragmentation, such as GALYFRE and DELFI.

A 2022 review identified 1,370 ongoing clinical trials using liquid biopsy for cancer detection in the United States (FDA) and Europe (EMA) [76]. These trials will invariably be investigating the use of mutations, methylation or fragmentation profiles of cfDNA. However, relatively few investigations have attempted to combine these distinct data types. A key limitation here is technical, where the cost of sequencing necessitates a trade-off between broad and shallow sequencing or deep, targeted sequencing. Mutation-based trials must necessarily perform targeted deep sequencing to stand a reasonable chance of detecting cancer-specific mutations in

their panels, whilst fragmentomics and methylation-based assays tend to require genome-wide information, at the expense of a shallow depth of coverage.

In this thesis, I focus on non-invasive detection of hepatocellular carcinoma (HCC) and pancreatic ductal adenocarcinoma (PDAC) from cfDNA liquid biopsy. These are two cancer types with particularly poor prognosis, mostly due to detection at an advanced disease stage [77, 78]. Non-invasive methods for early detection of PDAC and HCC are not available, which contributes to their late diagnosis. For decades, HCC detection has relied on liver ultrasound, combined with serum α-fetoprotein (AFP) measurements in blood [79]. However, these methods have low specificity and sensitivity [79, 80]. There is no blood test to detect or diagnose PDAC. Carbohydrate antigen 19-9 (CA19-9) is used for monitoring PDAC treatment and development, but its sensitivity and specificity are too low to diagnose or screen for PDAC [81]. Therefore, novel approaches for PDAC and HCC detection are urgently needed.

1.5 Deconvolution of cell-free DNA into its tissues of origin

Cell-free DNA can be extracted from a patient with a routine blood draw and sequenced affordably. If we can identify the relative contribution of each tissue to the cfDNA pool, we can infer the health status of each tissue and the patient overall. Whilst DNA is identical in all our cells (barring mutations), epigenetic marks are tissue-specific so can provide tissue of origin information. Each contributing tissue to the cell-free DNA pool has a tissue-specific methylation and hydroxymethylation signature. Epigenetic sequencing of this cfDNA pool captures the cfDNA genome-wide methylation signature. However, the observed sequencing signal from cfDNA is a mixture of tissue-specific signatures, and requires a suitable method for separating the observed mixture back into the source tissues that contributed to it. Assigning the proportional contributions of each tissue to an observed cell-free DNA methylation pattern is a signal separation problem known as tissue deconvolution.

Cell-free DNA tissue deconvolution has proved most successful using genome-wide methylation data as input. This combination is a particularly promising

avenue for multi-cancer early detection (MCED) tests, especially for cancers of unknown origin [60]. In 2022, a thorough evaluation was conducted on diagnostic performance of different cfDNA modalities, comparing mutation-based assays, cfDNA fragmentomics and methylation (n=2,800) [55]. They found that genome-wide cfDNA methylation was the single best predictor of cancer signal origin, and avoided the need for additional sequencing of patient blood cell background. Genome-wide methylation also demonstrated highest sensitivity for early stage disease, and was chosen as the most promising technology for MCED tests. The analysis of genome-wide methylation data for early cancer detection is the primary focus of this thesis.

1.6 Methylation sequencing methods

Before genome-wide methylation sequencing was affordable and widely available, profiling of CpG methylation state was performed by methylation arrays [82, 83, 84, 85]. The most widely used methylation array for such tasks is the Illumina (Infinium) 450K array [86]. Whilst cheap and convenient for targeted analysis, a significant limitation of these arrays is their genomic coverage: the Illumina 450K array covers less than 2% of CpG loci in the entire human genome [87, 88]. The use of pre-defined regions risks introducing bias into an analysis, by masking information from biomolecules that are not captured within the array. Furthermore, starting with the assumption that these are the maximally informative regions prevents discovery of new informative regions outside of the covered array, and can lead to an availability bias linking only the covered regions to observed phenotypes. Whole genome bisulfite sequencing (WGBS) has since become dramatically cheaper, and has become the *de facto* standard method for methylation sequencing [89]. Numerous studies employing genome-wide methylation sequencing have subsequently found that the majority of tissue-informative methylation markers lie outside of the regions covered in such arrays, and would have been missed with an array-based approach [90, 64, 91].

The output of methylation sequencing methods is a genomic DNA sequence where methylated cytosine can be distinguished from unmethylated cytosine. This is typically achieved by chemically treating extracted DNA before sequencing, so

that methylated and unmethylated cytosines are read separately in the resulting sequencing. Until recently, WGBS was the only available base-resolution methylation sequencing method, which requires harsh reaction conditions that can degrade up to 90% of sample DNA [92, 83]. This limits both the depth and breadth of coverage, and therefore its utility in clinical assays where the amount of patient DNA is limited. Reliability of downstream base calls is further hampered by low mapping quality due to loss in sequence complexity following conversion of unmethylated cytosine to thymine.

In 2019, a new method was developed in our lab by Liu *et al* [93], based on novel bisulfite-free chemistry: TET-assisted pyridine borane sequencing (TAPS). This method maintains sequence complexity, and avoids the harsh conditions of bisulfite treatment, thus reducing DNA degradation. The increased sensitivity of this non-destructive method enables the collection of high-confidence genome wide methylation patterns in clinical contexts where amount of DNA collected is small, such as liquid biopsies of cancer patients. Subsequent iterations of TAPS and a new chemistry known as CAPS [46] allow us to distinguish between 5-methylcytosine and 5-hydroxymethylcytosine. These new sequencing methods allows us to capture high quality genome sequence, methylome and fragmentation profile of cfDNA samples simultaneously, thus avoiding the technical bottlenecks or trade-offs discussed in the above trials.

In this thesis, I present the first analyses of this new sequencing method applied to cell-free DNA. I develop computational methods better suited to the improved resolution and data quality now available, aiming to make full use of the rich information provided. Specifically, I extend cfDNA tissue deconvolution to consider genome-wide tissue-informative markers, and develop a machine-learning method that may better scale to genome-wide input features. Secondly, I develop a classifier of cell-free DNA that incorporates the multiple data modalities captured in the TAPS assay. Due to improved data quality and resolution, this classifier can predict cancer origin at the level of individual cfDNA fragments, rather than using aggregated read counts across the genome.

1.7 Challenges of analysing biological sequence data

Our understanding of biology is fundamentally limited by our ability to measure a system without perturbing it. The development of technical and experimental tools to achieve this remains a major bottleneck in genomics and cell biology. Innovations in these experimental methods provide new forms of data: these allow us to examine biomolecules in greater resolution, which provides new information that can lead to new diagnostic tools.

Computational analysis methods evolve alongside these technical innovations, and generally lag behind their development. For example, when whole genome sequencing became possible, a new data type was effectively produced, containing vast new information but also bearing artefacts from the underlying sequencing chemistry employed to generate it. These data idiosyncracies meant new algorithms had to be developed to better process and analyse such high dimensional data, with its unique set of challenges, before meaningful insights could be extracted. The cost of genome sequencing has dramatically decreased in the last 20 years, which has led to a steep increase in the abundance of biological sequence data [94]. This represents a paradigm shift in the type of data available for biological studies, both in terms of volume and breadth across the genome. However, many existing computational methods are built for the previous data regime of targeted profiling of known dynamic sites in the genome and epigenome, such as exome sequencing or the use of arrays for methylation analysis [95, 85].

This trend holds true for cell-free DNA. The improvement in experimental technology, particularly genome-wide, high-depth sequencing has led to new information about the nature of cell-free DNA, and how it can inform tissue contributions and underlying physiology. For example, genome-wide cfDNA sequencing revealed that fragment length and distribution of cell-free DNA across the genome is highly non-random, and reflects nucleosome occupancy. This information led to new algorithms to predict tissue-specific gene expression using cfDNA fragmentation patterns [73, 74, 75].

We have only recently developed the technology to measure genome-wide methylation state of a tissue or blood sample, and our computational tools have not yet

caught up. There are several key challenges that arise with analysing this new genome-wide sequence data, such as the data presented in this thesis. They include:

1. The large breadth of these data sites, spanning the entire length of the genome. This is equivalent to a very high dimensional space where each data point is described with a very long feature vector.

2. The low sample size of most cohort based studies. Generally each sample is a patient, so data is scarce and expensive. Sample sizes on the order of tens or hundreds are common; a few large studies or national consortia can collect sample sizes in the order of thousands or above.

3. Lack of true biological replicates. It is difficult to obtain true biological replicates in many cases, as if samples come from human patients, each patient has a unique physiology and metabolism that is constantly changing through time and as a result of their environment. Often within-group variability is large, especially for cancer samples.

4. Myth of 'healthy controls'. Related to the above point, it is challenging to obtain a truly healthy control group where patient physiology can be assumed to be 'normal'.

5. Sequencing can involve non-random sampling of the underlying data, which introduces artefacts which must be properly accounted for in subsequent data analysis.

6. Precision of the analytical method should match resolution of the data generated. Computational methods can be excessively coarse and aggregate many data points into a single measure, when each individual data point can be informative. For example, only considering total read counts at genomic loci rather than each individual cfDNA fragment. Conversely, other methods assume a data resolution that is not warranted, such as over-interpreting local fluctuations in low-coverage sequencing that is more likely noise.

7. Difficulty of multimodal data integration. Biomedical datasets often have unique computational representations, which makes it difficult to integrate these datasets into a single analysis, combining orthogonal information to answer the same question. This is particularly of interest in this information scarce environment, where multiple modalities reinforcing one another can aid analyses.

1.8 Availability of public epigenetic data

There remain large gaps in our understanding of how cytosine methylation regulates cell phenotypes, and how this relationship changes in malignant neoplasms. Still less is known about the distinct role of hydroxymethylation, or the interplay of mC and hmC. These knowledge gaps are due to a paucity of high quality genome-wide methylation and hydroxymethylation data for diverse human cell types and tissues. Although there is an abundance of public omics data, the majority of methylation data is either array-based or exome profiling, so covers only a small fraction of the genome. Furthermore, public data tends to originate from a variety of sources, where samples from different backgrounds processed using different experimental conditions lead to batch effects between sources, making it harder to compare across samples. To rectify this, we constructed a genome-wide methylation and hydroxymethylation map of 22 tissues at base-pair resolution. This atlas also contains paired tumour data for 10 solid tumours. I develop a novel spatial segmentation algorithm to divide the genome into meaningful building blocks of mC and hmC, then identify tissue specific mC and hmC blocks which I validate with downstream gene expression analysis. This is the first reference atlas containing methylation and hydroxymethylation of healthy and tumour tissues, and represents a valuable community resource.

1.9 Machine learning methods for sequence data

Machine learning methods have garnered considerable interest in recent years, and have been successfully applied to many other domains. The term machine learning represents a vast array of methods, but the fundamental characteristic

is performing a task by iteratively minimising a cost function rather than using a set of rules or logic. The cost function links the input data to the success criteria, thus allowing the model to learn from the data. Most successful methods in recent years have used automated feature extraction on the input data, rather than manually extracted features. By linking the cost function to the weighting of the input data, this approach allows the priority of features in the input data that are informative for the task at hand, and disregard of uninformative features. Deep neural networks are a model architecture that have proven to be particularly good at automated feature extraction. These models place weights on all input data points which are easily adjustable through backpropagation, and combine this with non-linear transformations of input data across multiple layers. The extensibility of this approach and the ability to fine-tune many millions of individual parameters allows these models to act as universal function approximators [96].

Machine learning methods have also been widely applied on biomedical datasets, and have been particularly useful for analysing medical images such as CT scans [97], and electronic health record data [98]. To date, there have been fewer genuinely useful applications of machine learning for biological sequence data. This is largely due to reliance on supervised learning methods, which require large amounts of labelled data. Most real world clinical datasets have low sample size and are high dimensional, often with many more features than samples. There are relatively few applications of deep learning methods to methylation data. Notable exceptions include DeepCpG and MRCNN to predict CpG methylation status using DNA sequence [99, 100], and DISMIR, a recent read-classification method that incorporates methylation [101]. A key bottleneck here has been data scarcity: not enough genome-wide methylation data has been publicly available to warrant training large machine-learning models. The introduction of our tissue and tumour atlas resource will help change this.

Patient data will remain scarce and expensive, but omics data is increasingly abundant and cheap. One way to avoid sample size limitations is to build models where each data point is not an individual, but a different biological entity such as a gene or a protein, effectively modelling in a different domain where training data is abundant. This is a method I use in my thesis, for the fragment-level classifier model. Machine learning methods have been successfully applied to problems

such as protein structure prediction from amino acid sequence [102], to predicting gene transcription factor binding sites and gene expression from DNA sequence [103, 104]. These approaches use RNNs, LSTMs and Transformer neural network architectures, which are specially adapted to handle sequential data [105, 106, 107]. By treating each gene or protein as an individual data point, we now have many thousands of examples to train large neural network models that can act as excellent feature extractors across the genome. Most notably, the Enformer[104] model uses a Transformer neural network architecture to scan across 200kb of DNA surrounding each gene, and simultaneously predicts gene expression across multiple tissue types using DNA sequence alone. Accurate models learn an effective mapping from the data input space to the desired output, and which input features are most informative for this task. For models trained on biological sequences, probing this learnt mapping can provide new information about the original biological molecules. For example, ablation experiments in the Enformer model can predict variant effects on gene expression, and feature attribution methods reveal regulatory sequence motifs.

A second approach that circumvents the small training data problem in healthcare is training a model on an adjacent task where data is abundant, then applying the trained model to the desired task by transfer learning. This strategy has proven effective for building large language models (LLMs) on text data. Here, a model is trained on the self-supervised task of next word prediction with large amounts of text, and learn to generate excellent internal representations of the nature of the input data. These internal representations can then be harnessed for fine-tuned on specific tasks such as sentiment analysis, entity recognition or text generation by transfer learning [108, 109].

Similar approaches have been developed for biological sequence data. Protein language models, large self-supervised models trained on millions of amino acid sequences, are becoming useful for predicting secondary protein structure and for designing novel proteins with desirable characteristics [110, 111]. In genomics, large models such as Enformer and Borzoi [112] are becoming accurate predictors of gene expression across tissue types using DNA sequence alone. DNABERT-2 and HyenaDNA [113] are both accurate predictors of promoter and enhancer positions across the genome using DNA sequence alone, and HyenaDNA can additionally

predict location of epigenetic marks [114]. The use of transfer learning on adjacent tasks to preserve patient data is a strategy I employ in the cfDNA deconvolution and classification models developed here.

1.10 List of publications

This thesis includes the following published work, in which I was first author or joint first author:

Cell-free DNA TAPS provides multimodal information for early cancer detection. Science Advances, 2021

Multimodal Cell-Free DNA Embeddings are Informative for Early Cancer Detection. NeurIPS 2022 Workshop: Learning Meaningful Representations of Life

Deconvolution of cell-free DNA in cancer liquid biopsy using a deep autoencoder. BCB '23: Proceedings of the 14th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics

Predicting gene expression using genome-wide patterns of cytosine methylation and hydroxymethylation. Machine Learning in Computational Biology Conference, 2023 (poster presentation)

Chapter 2

Background

2.1 The nature of methylation sequencing data

Whole genome methylation sequence data at CpG-level resolution consists of aligned read counts at 28 million individual CpG sites. At each CpG site i, we have a count C_i of reads containing methylated cytosine, a count U_i of reads containing unmethylated cytosine and a count T_i of total reads at that locus, such that $T_i = C_i + U_i$. Methylation ratio is calculated as:

$$\text{CpG methylation ratio} = \frac{C_i}{T_i} \tag{2.1}$$

The mean methylation level across a region of length n CpG sites is simply the mean of all site CpG methylation values, calculated as:

$$\text{Region methylation ratio} = \frac{1}{n}\sum_{i=1}^{n}\frac{C_i}{T_i} \tag{2.2}$$

An alternative calculation for region level methylation is to separately sum methylated and total read counts calculating their ratio, as below. This weighted methylation ratio is useful when individual CpG coverage is low or uneven, so incorporating methylation ratio at individual CpG sites is unreliable and may introduce bias, whereas taking read counts summed across an entire region is more reliable.

$$\text{Weighted region methylation ratio} = \frac{\sum_{i=1}^{n} C_i}{\sum_{i=1}^{n} T_i} \tag{2.3}$$

The nature of methylation data and how it was generated must be taken into account for all subsequent analysis. It should be recalled that the observed methylation ratio at any given locus is just an imperfect representation of the true underlying methylation state. Deep sequencing is still prohibitively expensive, so shallow sequencing (<5x) is often used, which can lead to large variation in depth of coverage, with order of magnitude changes in read coverage across genomic loci. Can we be sure that a CpG covered by just four reads does indeed have a methylation ratio of 0.25, or is this a result of a random sampling error? One must also consider the sample from which the DNA was obtained. If taken from a tissue biopsy, this is likely to contain lots of individual cell types with distinct methylation states at any given locus: the methylation state we observe is the mixture of these individual cell types. Similarly for cell-free DNA methylation sequencing, we observe only a sample of total circulating cfDNA fragments, and make assumptions that this sample is representative of the population. The statistical challenges faced when analysing methylation data are illustrated below by considering the common task of identifying differentially methylated regions.

2.2 Identifying differentially methylated regions

Differentially methylated regions (DMRs) are genomic loci exhibiting differential methylation between two sample groups. Methylation is known to play a central role in tissue differentiation in both healthy and malignant cell fates, and is highly tissue specific. The information content of methylation is conveyed through increases or decreases in the methylation level of a genomic locus. Differential methylation is the epigenetic counterpart to differential expression, however differential methylation occurs upstream of gene expression, and is to an extent the causative change affecting downstream gene expression. A natural first step in any analysis of methylation data is therefore to identify tissue specific methylation markers, and it is used several times throughout this thesis. This involves searching through the 28 million CpGs for tissue informative regions. From a statistical perspective, we start with a very high dimensional vector space representation of tissue methylation state, and need to perform feature selection to enrich for informative signal and reduce overall noise. The high breadth of genome-wide

methylation sequences makes tasks like DMR identification a statistical hypothesis testing task similar to GWAS, where some assumptions must be made regarding the population distribution of the data which is being sampled, and care must be taken to correctly handle multicollinearity and multiple test corrections.

DMR finding algorithms generally consist of two distinct steps.

1. Grouping the methylome into distinct regions

2. Some statistical test to compare methylation in target group versus a background, performed over each independently.

This is the usual ordering of these steps, but in some methods these steps are reversed: we first identify many individual differentially methylated cytosines, then group these together into contiguous segments of differential methylation. In both cases, DMR calling has both a spatial component horizontally along the genome, and a vertical effect size component.

Three statistical challenges present themselves when analysing methylation data.

1. There is large variability in methylation states within sample groups. Each sample often comes from a separate person with unique physiology and genetics, so this variation is biologically meaningful and cannot be accurately summarised by pooling or taking mean values

2. Shallow methylation sequencing technologies lead to large variation in depth of coverage, with order of magnitude changes in read coverage across genomic loci. Regions with shallow coverage (< 5x) are prone to random sampling error, leading to differential methylation values that are not representative.

3. DNA methylation demonstrates local spatial correlation. The correct statistical model should take into account the scale at which this spatial correlation exists, such that nearby methylation values can be used to reduce uncertainty of model performance in low coverage regions, whilst maintaining the ability to detect true boundaries in genomic elements, such as the border of a transcription factor binding site [68]

Given the relatively recent arrival of whole genome methylation sequencing, consensus has not been reached on the correct statistical approach for modelling genome-wide methylation and identifying DMRs. Table 2.1 reviews many of the key published DMR finding algorithms. These are a sample of the most often used and cited methods in the literature, but many more programs are currently available, each with a bespoke software implementation and different data formatting requirements. However, the statistical underpinnings of these methods overlap significantly, and can be usefully categorised based on how they deal with these three challenges.

Method name	**Publication Date**	**Genomic Region**	**Model Assumption**	**Statistical Test**
BSmooth [115]	2012	Merging adjacent DMLs	Binomial distribution	T-test
methylKit	2012	Merging adjacent DMLs	Logistic regression	Logistic Regression
RADMeth [116]	2014	Merging adjacent DMLs	Beta-binomial	Log-likelihood ratio test
MethylSig [117]	2014	Merging adjacent DMLs	Beta-binomial	Log-likelihood ratio test
DSS	2015	Individual DMLs	Bayesian hierarchical model	Wald test
Metilene [118]	2016	Define segments of homogeneous methylation	Non-parametric	2D Kolmogorov–Smirnov
DMRFinder [119]	2017	Clustering neighbouring CpGs	Beta-binomial	Wald test
MethAtlas [85]	2018	Fixed windows	None	Delta threshold
DMRcate [120]	2021	Merging adjacent DMLs	Beta-binomial	Modified Limma
CelFiE [121]	2021	Individual CpGs which are then merged	None	Delta threshold
DISMIR [101]	2021	500bp windows across genome	None	Delta threshold

Table 2.1: Summary of methods for differential methylation analysis

The simplest method of comparing whole genome methylation simply compares

the group-wise mean at each CpG site or region of interest, using Fisher's Exact Test or similar. The limitation of not taking within-class variability into account has been illustrated numerous times in the literature [95, 122]. The scaling parameter of T-tests allow sample means to be compared whilst scaling by within-group standard deviation, yet their reliance on a point estimate of the true methylation distribution fails to capture the actual variance in methylation values, which arises from coverage discrepancies as well as biological noise. Furthermore, with small numbers of biological replicates, the estimate of within-group variance is unstable, and the asymptotic assumption of a t-distribution may be invalid [123].

2.3 Modelling assumptions for methylation data

More recently, DMR finding methods have modelled target loci using the beta binomial probability distribution. This distribution has also been co-opted more widely as a good underlying data generation model for CpG methylation, in tasks such as cfDNA simulation and cfDNA classification using probabilistic inference [121, 124]. I use this distribution for data simulation of cfDNA for use as training data. The principles are as follows.

The binomial distribution is a natural choice to represent the number of methylated reads from the total at a given locus, which is analogous to k successes in n trials and is dependent on the methylation proportion P. As mentioned, for any given locus this methylation proportion can vary significantly between biological replicates, so a beta distribution is used to describe the distribution of parameter P across the group. The beta distribution is generally reparameterised in terms of a CpG specific mean methylation ratio and a dispersion parameter [122]. By describing the values we can measure - methylated read count, total read count, methylation ratio - as random variables with probability distributions, we can find the likelihood of the values we observe. Then, by relating the probability of the observed data to the parameter we want to estimate - in this case θ - we can make a best guess using maximum likelihood estimation: finding the parameter value (θ) that maximises the likelihood of observing the data we have. The steps are described in detail below:

1. Observed methylated read count, at a given CpG, can be modelled with a binomial distribution. Any given CpG site on a cfDNA fragment is either methylated or unmethylated, and the probability of each outcome (1 or 0) is determined by a prior probability p [1] . This prior is approximated by the methylation ratio at CpG k, x_k

$$m_k \sim Binomial(n_k, x_k) \tag{2.4}$$

2. This methylation ratio can be modelled with a beta distribution. At a given CpG, reference samples within a class (eg all healthy samples) have a range of MR values, which are used to fit the parameters of a beta distribution (using MLE/method of moments). Importantly, because MR represents the probability of a read being methylated, this distribution of MR values can be used as the prior probability of methylation, p, in 2.4. Thus, at each marker, all MR values within a class are represented by a single beta distribution.

$$x_k \sim Beta(\alpha_k, \beta_k) \tag{2.5}$$

3. Considering the likelihood of the observed methylated read counts, m_k, is a binomial distribution that depends on x_k, as in 2.4 above, we can substitute our expression for x_k in 2.5 into 2.4.

   Thus, the likelihood of the observed methylated read counts at CpG k is given by a compound beta-Binomial distribution:

$$m_k \sim Binomial(n_k, x_k)Beta(\alpha_k, \beta_k) \tag{2.6}$$

[1] A useful analogy is a coin toss where there are two outcomes (H or T): the probability of heads is dictated by a prior based on the coin (P=0.5 for a fair coin).

2.4 Context is key: defining the background in DMR-finding algorithms

An important question to ask when searching for differentially methylated regions is: differentially methylated compared to what? The choice of the background group we are comparing to is key, and greatly influences the type of statistical methods employed, and the actual DMRs we are likely to find. A differentially methylated region is not just a stable list of CpGs, but a reflection of the samples used to generate it. The vast majority of DMR finding algorithms have been defined for pairwise comparisons in a case versus control scenario, for example methylation changes in a healthy cell type versus a cancer cell type [117, 118, 120]. The DMRs found with this pairwise method will be good at distinguishing these two cell types, but may not be able to distinguish these cell types in a broader context, as illustrated in Figure 2.1. This region would therefore be a DMR in the first context but not in the second context.

This has practical implications for our setting, when trying to find tissue specific DMRs in the tissue atlas. We may find a hypermethylated region that can distinguish colon tissues from most other tissues and blood cell types, but developmentally linked tissues such as esophagus and stomach exhibit a similar hypermethylation state at that locus. Equally, we may find a tumour DMR that clearly distinguishes it from its healthy tissue, but all blood and immune cells demonstrate a similar methylation state. This emphasises the importance of having a broad range of tissue types in the atlas to form the background to compare against: more tissue types in a reference decreases potential for introducing bias. Ultimately we want to use these DMRs for improved tissue deconvolution of cell-free DNA, so we want DMRs to be as tissue informative as possible, and not an artefact of the specific tissue types contained in the reference.

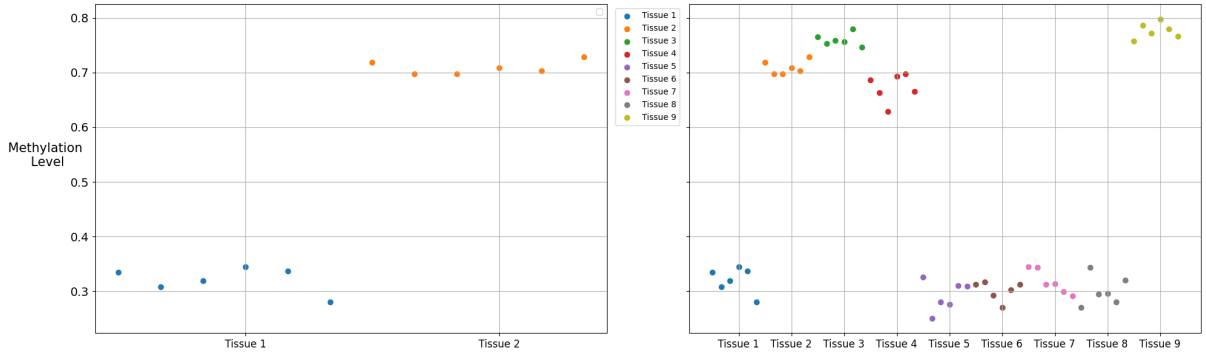


Figure 2.1: An example of finding DMRs on synthetic data

2.5 Defining DNA methylation signatures

DNA methylation signatures, known as "episignatures" are beginning to emerge in connection with various rare disorders [125]. These episignatures are loosely defined as a pattern of differential methylation across multiple genomic loci specifically associated with a particular phenotype. A classic example is Horvath's epigenetic clock signature [126]: a collection of 346 CpGs selected and weighted using ElasticNet regression, whose methylation status is frequently used to determine human epigenetic age. More recently, DNA methylation signatures have been detected for pathologies such as Floating-Harbor syndrome [127], chronic low-grade inflammation [128], Sotos syndrome [125] and even environmental exposure to cigarette smoke [129]. These signatures have generally been defined by regression methods to assign weights, and more recently machine learning methods [130]. These episignatures differ from DMRs in that, rather than just representing a single region, they represent a collection of regions, which when combined with appropriate weightings can distinguish a tissue or phenotype of interest.

2.6 Measuring tissue specificity of candidate DMRs

At several points in this thesis it was necessary to measure tissue-specificity of candidate DMRs, to identify maximally informative DMRs before proceeding with downstream analysis based off these regions. To achieve this, I identified statistics originally devised for economics but which have been used to quantify tissue

specificity of gene expression, and adapted these for methylation data. These were Gini impurity coefficient [131] and Shannon entropy specificity [132]. These scores were defined as follows:

Gini coefficient

The Gini coefficient is a measure of inequality across groups. A high Gini coefficient here would therefore indicate that few tissue samples have very distinct methylation levels compared to the rest, thus indicating a more tissue specific DMR.

For a set of methylation values, $X = \{x_1, x_2, \ldots, x_n\}$, the Gini coefficient, G, is calculated as:

$$G = \frac{\sum_{i=1}^{n}(2i - n - 1)x_{(i)}}{n\sum_{i=1}^{n} x_{(i)}} \tag{2.7}$$

where $x_{(i)}$ is the i-th smallest value in the sorted list of methylation levels, and n is the total number of tissues.

Shannon Entropy Specificity

Shannon entropy specificity measures the diversity of methylation distribution across different tissues. For a given set of methylation values, X, the Shannon entropy specificity is calculated as follows:

$$H(X) = -\sum_{i=1}^{n} p_i \log_2(p_i) \tag{2.8}$$

$$\text{Shannon Entropy Specificity} = \log_2(n) - H(X) \tag{2.9}$$

where $p_i = \frac{x_i}{\sum_{j=1}^{n} x_j}$ is the proportion of methylation in the i-th tissue, and n is the number of tissues.

To further validate these tissue specificity metrics, I also used statistics to measure the extent of tissue specific clustering, after intersecting tissue samples on a set of candidate tissue markers. I quantified overall quality of clustering using the unsupervised (meaning tissue group labels not provided) Silhouette Score metric [133], then used two supervised metrics to quantify the extent to which tissue samples clustered into their tissue groups. These were adjusted Rand index [134] and normalised mutual information score [135]. These metrics are summarised briefly as follows:

Silhouette Score

The Silhouette Score is an unsupervised metric used to assess the quality of clustering. For each sample, it calculates the difference between the average distance to samples in the same cluster and the average distance to samples in the nearest cluster. The score ranges from -1 (poor clustering) to 1 (highly dense clustering). It is defined as:

$$S = \frac{b - a}{\max(a, b)} \tag{2.10}$$

where a is the mean distance to the points in the same cluster, and b is the mean distance to the points in the next nearest cluster.

Supervised evaluation of assigned clusters

Normalized Mutual Information is a supervised metric used to measure the quality of clustering by comparing the cluster assignments to the true labels. It quantifies the mutual information between the clustering and the true labels, normalized to account for the size of the clusters. It is defined as:

$$\text{NMI}(C, T) = \frac{2 \times I(C; T)}{H(C) + H(T)} \tag{2.11}$$

where $I(C; T)$ is the mutual information between the clustering C and the true labels T, and $H(C)$ and $H(T)$ are the entropies of C and T, respectively.

The Adjusted Rand Index is another supervised metric used for evaluating clustering. It is defined as the number of agreement pairs divided by the total number of pairs, adjusted for chance. It ranges from -1 (random labeling) to 1 (perfect match). The ARI is defined in [134].

Chapter 3

Tissue deconvolution of cell-free DNA methylation data

3.1 Abstract

Tissue deconvolution of cell-free DNA methylation data is a promising method for non-invasive early cancer detection. However, available methods to perform this deconvolution have largely been built for and benchmarked on methylation array data. These arrays only capture around 2% of genomic CpGs, and so risk excluding tissue-informative markers and biasing the assigned tissue contribution. To rectify this, I construct a comprehensive tissue atlas of public genome-wide methylation data. I use this reference atlas to define tissue-specific differentially methylated regions, using a new method I develop to maximise capture of tissue informative regions. The resulting tissue markers are used as a reference for deconvolution of cell-free DNA methylation data from the TAPS cfDNA sequencing pilot study. I identify cancer-specific tissue contributions, such as a signficantly increased liver-tumour contribution in liver cancer patients, and demonstrate this method can be effective for distinguishing cancer patients from non-cancer controls. Seeking to confirm this deconvolution result, I compare NNLS deconvolution results to a newer deconvolution method that employs maximum-likelihood inference, known as CelFiE.

Finally, I develop a deep learning based cfDNA deconvolution method, demonstrate its improved performance on simulated cfDNA, then apply this novel method

on patient cfDNA samples. I show that trained model predictions of tissue contribution agree with other widely used approaches, and that it too can be used to distinguish cancer cfDNA samples from non-cancer controls.

3.2 Introduction

Cell-free DNA (cfDNA) is released from tissues and blood cells during apoptosis, and circulates freely in peripheral blood. The relative concentrations of DNA released from each tissue can tell us about the underlying health and physiological status of that tissue. For example, during vigorous exercise, muscle cells undergo high turnover, and may result in elevated cfDNA from skeletal muscle cells detected in the blood afterwards. Similarly, when a malignant tumour arises in a tissue or blood cell, cancer-specific changes to cell physiology such as increased cell proliferation can be reflected in cell-type specific cfDNA shedding. The tumour itself also sheds DNA into the cfDNA pool, called circulating tumour DNA (ctDNA). Epigenetic sequencing of this cfDNA pool captures the cfDNA genome-wide methylation signature. However, the observed sequencing signal from cfDNA is a mixture of tissue-specific signatures, and requires a suitable method for separating the observed mixture back into the source tissues that contributed to it.

Assigning the proportional contributions of each tissue to an observed cell-free DNA methylation pattern is a signal deconvolution problem, where the raw input signal is a vector of cfDNA methylation across the 29 million CpGs in the human genome. Cell-free DNA deconvolution is a biomolecular example of a more general problem in signal processing known as blind source separation. This arises in audio processing when trying to separate a musical recording into contributions from individual instruments, and in text processing when identifying underlying topics and their frequencies in documents of text, using methods such as Latent Dirichlet Allocation [136]. Closer to home, an analogous problem exists for gene expression data: trying to decompose bulk RNA-sequencing data into contributions from individual cell types [137, 138]. This signal deconvolution problem is summarised for cell-free DNA in Figure 3.1.

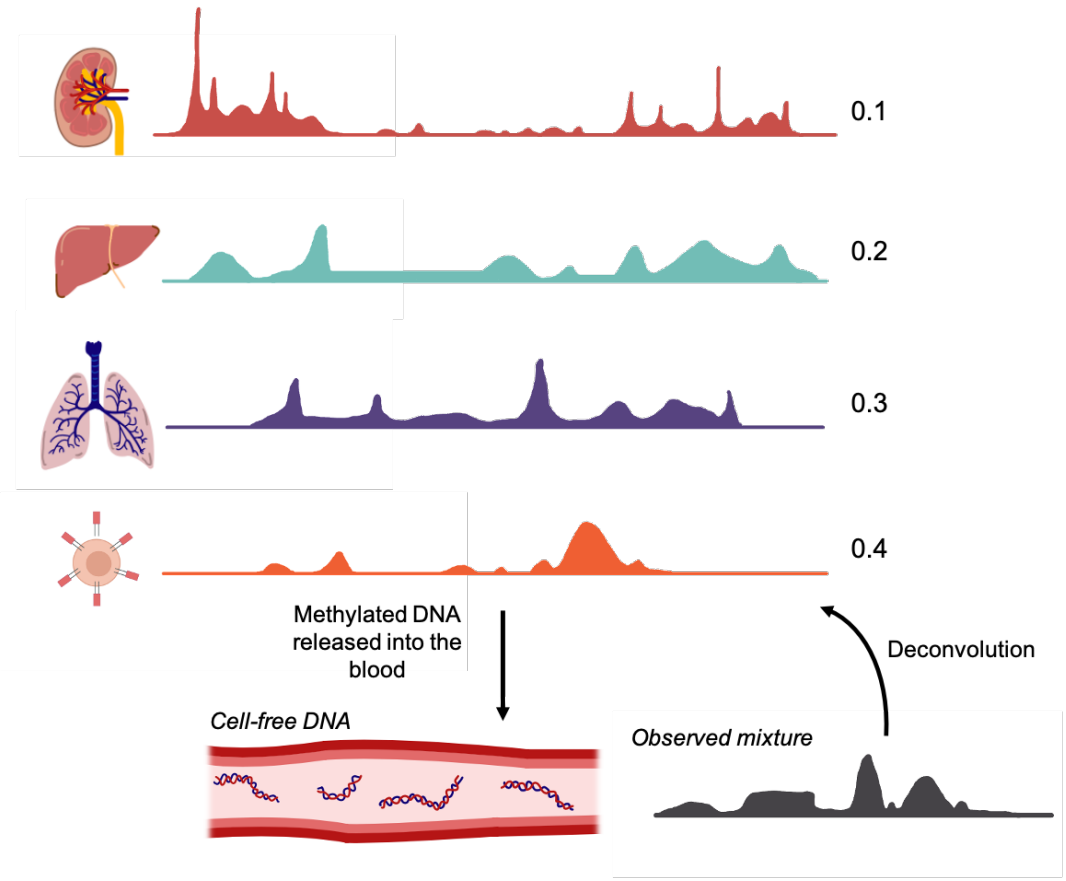


Figure 3.1: Schematic illustration of the cfDNA tissue deconvolution problem. Each tissue sheds DNA into the blood containing a unique methylation signature, which then circulates as cell-free DNA. The resulting cfDNA can be extracted from blood and sequenced non-invasively, but the observed mixture must be separated into its constituent components.

Deconvolution of cfDNA methylation can be formalised as follows. We have a set of k tissues (and blood cell types), each with a distinct methylation signature defined over m CpGs, all contributing DNA into the cfDNA pool. Each tissue has its own specific rate of shedding into the cfDNA pool, which is defined with frequency matrix F. The methylation pattern in the cfDNA pool which we observe after sequencing is therefore a product of all contributing tissue signatures, and their individual frequencies. The task of cfDNA deconvolution is to obtain the

matrix of proportions A, which contains the relative proportion of each underlying tissue in each sample of interest. This is illustrated in Figure 3.2.

Deconvolution approaches fall into two categories based on how we treat the matrix T in our schematic: reference-based, and reference-free deconvolution. Reference-based approaches use a reference of epigenetic signatures T, which are combined with the observed cfDNA mixture D to solve for the matrix of proportions A. Reference-free methods instead jointly estimate the underlying signatures T contained in the observed mixture, as well as the proportional contributions A of each signature. In both reference-free and reference-based deconvolution the k tissues originally contributing to the cfDNA pool may not be equal to the p tissues captured in the reference, T.

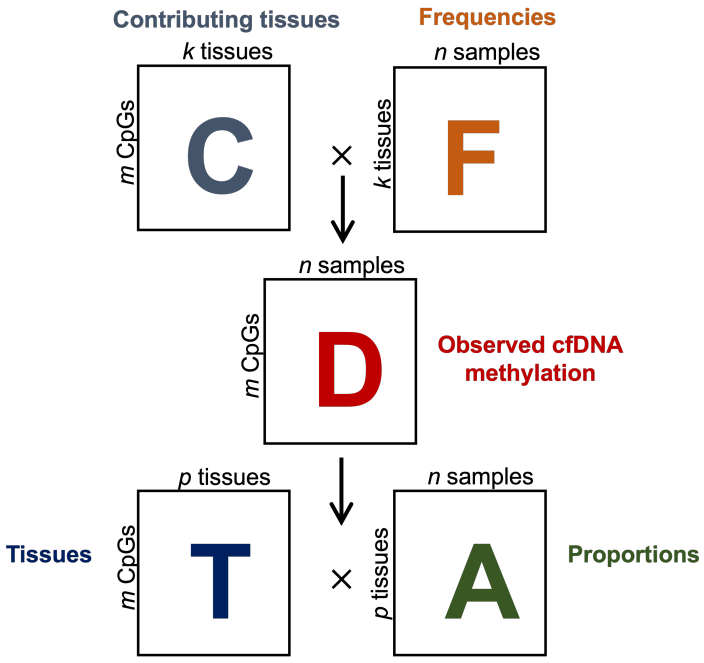


Figure 3.2: Formalisation of the cfDNA tissue deconvolution problem in matrix form. k tissues shed cfDNA into the blood at tissue-specific rates captured in matrix F, forming a cfDNA pool which is the observed mixture D. Deconvolution methods separate this observed mixture into a matrix of proportions A and a reference T.

Tissue deconvolution of cfDNA faces several unique challenges. Firstly, all

tissues can release cfDNA into the blood, but only a relatively small proportion of cfDNA molecules originate from solid tissues: the vast majority (80-90%) of cfDNA contributions are from blood cells. Changes in cfDNA contribution from solid tissues undergoing physiological changes of interest are therefore hard to detect, amidst the background of blood cell cfDNA. Indeed, there are other determinants of detectable cfDNA concentration that are not directly related to tissue health, which is generally the underlying phenotype we hope to measure. These include physical proximity of tissues to blood vessels, stability and rate of clearance of cfDNA in the blood, and metabolic activity of individuals through time.

Despite the relatively low contribution of solid tissues and other cell types, it is currently assumed that all tissues can release detectable amounts of cfDNA, which highlights the second challenge: the need for a reference dataset that encompasses epigenetic signatures of all possible contributing tissues and cell types. Reference-based cfDNA deconvolution methods infer the optimal proportions of each tissue in the reference to the observed cfDNA mixture, but are inherently limited by the tissues contained in the reference. We do not yet have epigenetic profiles of all the cell types in the human body that could contribute to the cfDNA pool, so any reference is necessarily limited and introduces an availability bias.

The diagnostic potential of cfDNA means many attempts have been made to deconvolute cfDNA, but the range of computational methods employed are limited. Most published approaches have used one of two reference-based cfDNA deconvolution algorithms: non-negative least squares regression (NNLS), or likelihood-based expectation-maximisation.

Non-negative least squares regression was first used for cfDNA methylation deconvolution by Moss *et al* in 2018 [85], and is still widely used in recent studies [64, 91]. This is a constrained minimisation problem, where given an m x p reference atlas, T, describing methylation ratios of p tissues at m positions, and a m x n matrix of observed methylation across n samples or individuals, we find a matrix of coefficients, A, such that:

$$\begin{aligned} &\text{argmin}\, ||TA - D||^2 \\ &\text{subject to}\, A \geq 0 \end{aligned} \tag{3.1}$$

The coefficient matrix A represents the tissue contribution here. Quadratic programming can find a locally-optimal solution to this problem, but does not guarantee a global minimum. Importantly, this method gives no indication of the uncertainty of each tissue assignment, or the amount of cfDNA of unknown origin in a sample. Consider the case where 40% of the cfDNA contribution is predicted to come from tissue x: if we then remove tissue x from the reference atlas, this method will confidently attribute this methylation signature to other tissues with no indication of a reduction in assignment certainty. Despite these limitations, the majority of large studies have so far favoured this approach, due to a lack of alternative options.

A recent attempt to overcome the limitation of incomplete reference data is CelFiE [121]. This method takes a more rigorous likelihood-based framework, using total and methylated read counts of both reference matrix and observed cfDNA to parameterise a beta-binomial probability distribution (as described in Background chapter), which it then optimises using expectation-maximisation. CelFiE incorporates the contributing cell types k as a latent variable in the model objective, which allows it to include contributions from unknown cell types which are missing in the reference. This algorithm has recently been extended to work incorporate fragment-level information rather than read count averages at each locus, which may improve its sensitivity to trace contributions [139]. By taking a principled approach of modelling observed methylation data with commonly used distributions, these likelihood based approaches are also more robust to noise that is often present in low-coverage sequencing data.

Reference-free methods circumvent the *a priori* assumptions of reference-based deconvolution by removing the need for a reference altogether. These approaches jointly estimate the underlying signatures and the proportional contributions simultaneously, which is much more challenging statistically due to to the reduced input information. The most commonly used reference-free approaches are variants of non-negative matrix factorisation, which iteratively minimises the cost functions associated with estimating matrices A and T [140, 141]. These algorithms infer a matrix of proportions, and latent methylation components (LMCs) corresponding to the 'true' underlying signatures contained in a mixture.

A key limitation with these methods is we must still stipulate the number of expected underlying signatures, and there is no reliable way of knowing this number in advance, or indeed finding the optimal number [142]. These algorithms are generally computationally costly, and computational cost scales with the proposed number of underlying LMCs. We cannot therefore propose the number of underlying LMCs to be the total number of cell types in the human body, so for practical purposes we are limited to 20-30 underlying signatures. Computational cost also scales with number of genomic loci in the input data, necessitating a large reduction in the possible input feature space, which limits their utility on genome-wide data. For these reasons, reference-free methods of cfDNA deconvolution have not been widely adopted on clinical data, despite their theoretical advantages.

The choice of genomic loci over which cfDNA deconvolution is performed can dramatically impact the inferred contribution. To distinguish tissue-specific contributions, we must use genomic regions that are tissue informative. All current tissue deconvolution algorithms require some form of feature selection before deconvolution, as they lack the statistical and often computational power to be run on all 28 million CpGs. There is therefore a compromise to be reached in providing a broader or narrower range of genomic features as model input. Up to a certain point, considering a wider range of genomic features increases the tissue-specific information captured. However, different statistical models vary in their capacity to make use of this information and actually translate it to more accurate deconvolution. Including all genome-wide CpGs is likely to include lots of superfluous noise that is not tissue informative. Ultimately, we want our choice of genomic features to be maximally tissue informative and minimally noisy.

Previous deconvolution methods have largely relied on methylation arrays, especially the Illumina 450k array (Infinium HumanMethylation450 BeadChip) [83, 84, 124]. At the time of this work there were no available tissue reference datasets containing genome-wide methylation data. As a result the genomic loci considered for deconvolution were limited to those contained within these arrays, which amounts to around 2% of genomic CpGs. These arrays may therefore not contain the maximally tissue informative CpGs for a particular task or context, and as discussed risk introducing bias into deconvolution experiments.

In this chapter I take a genome-wide approach to identifying tissue informative regions for subsequent cfDNA deconvolution. I first collate a reference atlas of genome-wide tissue methylation signatures from public whole genome bisulfite sequencing (WGBS) data. I develop a new method for identifying maximally informative tissue-specific differentially methylated regions within this tissue atlas, then use these regions for NNLS deconvolution of patient cfDNA. I demonstrate this method can be effective for distinguishing cancer patients from non-cancer controls, then compare NNLS deconvolution to other approaches such as CelFiE.

All deconvolution methods mentioned so far employ an unsupervised approach, whereby a tissue contribution is assigned without first training a model for this task. Supervised machine learning has proven to be a very powerful system for efficiently learning from large amounts of data to perform a specialised task highly accurately. In particular, deep neural networks are universal function approximators and have outperformed linear regression methods on non-linear objectives. In the final part of this chapter, I develop a deep learning based cfDNA deconvolution method and demonstrate its performance on simulated and real cfDNA samples. I show that model predictions of tissue contribution agree with other currently used approaches, and that it too can be used to distinguish cancer cfDNA samples from non-cancer controls.

3.3 Assembling a genome-wide tissue methylation reference atlas

For reference-based cfDNA deconvolution, to minimise bias introduced by the reference, we need to maximise the number of tissues captured in the reference, and the breadth of genomic loci covered with high confidence read counts. To ensure that each tissue signature is representative, it is helpful to collect many biological replicates from each tissue, and use these replicates to remove anomalous samples. I sought to collate the largest tissue methylation reference atlas yet assembled, by downloading all publicly available whole genome bisulfite sequencing (WGBS) data for healthy tissues. The stipulations for inclusions of tissue methylation datasets in our reference atlas were as follows:

1. Must contain genome-wide information across all chromosomes.
2. Must be at CpG-level resolution, not averaged across regions.
3. Must be healthy tissue, sourced from a healthy, adult donor.
4. Must contain total as well as methylated read counts at each genomic position, to ascertain coverage.
5. Must contain linked metadata describing how it was sourced, and basic demography and health status of patient donor.

I downloaded 208 tissue WGBS samples, from 6 public data consortiums: ENCODE, CREST, Roadmap, Blueprint, DEEP and CEEHRC. Six additional liver and liver tumour WGBS samples were obtained from personal correspondence with Li *et al* [124]. After filtering diseased, sex-specific, and low-coverage samples, we retained 144 healthy, adult tissue samples, grouped into 32 physiologically distinct tissue groups. Of these 144 samples, 133 were already aligned to hg38; the remaining 11 samples were converted from hg19 to hg38 using the UCSC hgLiftOver tool [143]. Tissue data sources are summarised in Figure 3.3. Full details on all tissue samples included in this atlas are provided in Table S5 of the paper Supplementary Information [144].

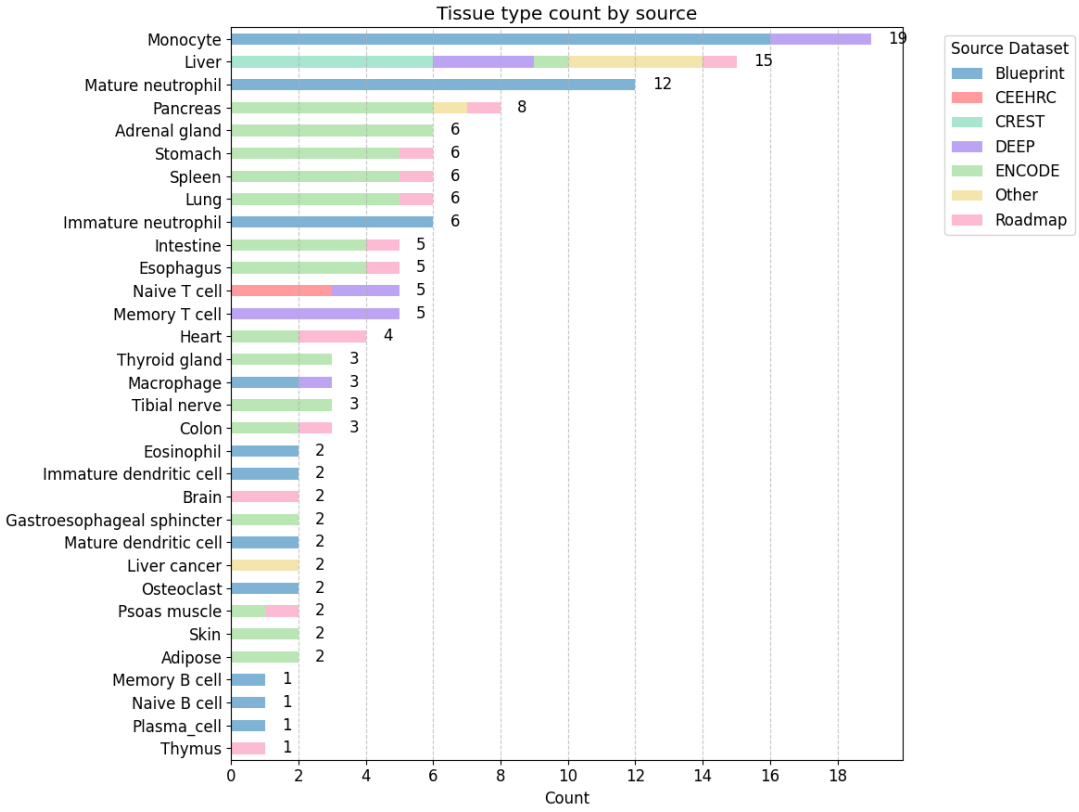


Figure 3.3: The number of samples for each tissue in our final reference dataset after filtering samples, split by source dataset. Liver and liver tumour samples marked 'Other' were obtained from personal correspondence with Li *et al.*

3.3.1 Unsupervised clustering of tissue samples

Due to inconsistent and often low coverage of some tissue samples in the atlas, I elected to use mean methylation across functional regions as the input for all downstream analysis. I extracted genomic location files for enhancers, transcription factor binding sites, promoters, CpG Islands and open chromatin regions from the Ensembl Regulatory Build [145]. These are determined by a Hidden Markov Model based on ChIP-seq measured histone modifications, for example Enhancers are characterised as Dnase I hypersensitive sites enriched in H3K27ac signal. For each feature type, I calculated mean methylation state over each feature, then clustered tissue samples with hierarchical clustering. I found enhancers demonstrated best tissue specific clustering, so used enhancers for all subsequent analysis. This

decision was supported by an emerging literature highlighting the important role enhancers play in tissue specific regulation of gene expression [146, 147, 148, 149]. Figure 3.4 shows a t-SNE embedding calculated over all enhancer regions, of all tissue samples in the assembled reference atlas [150].

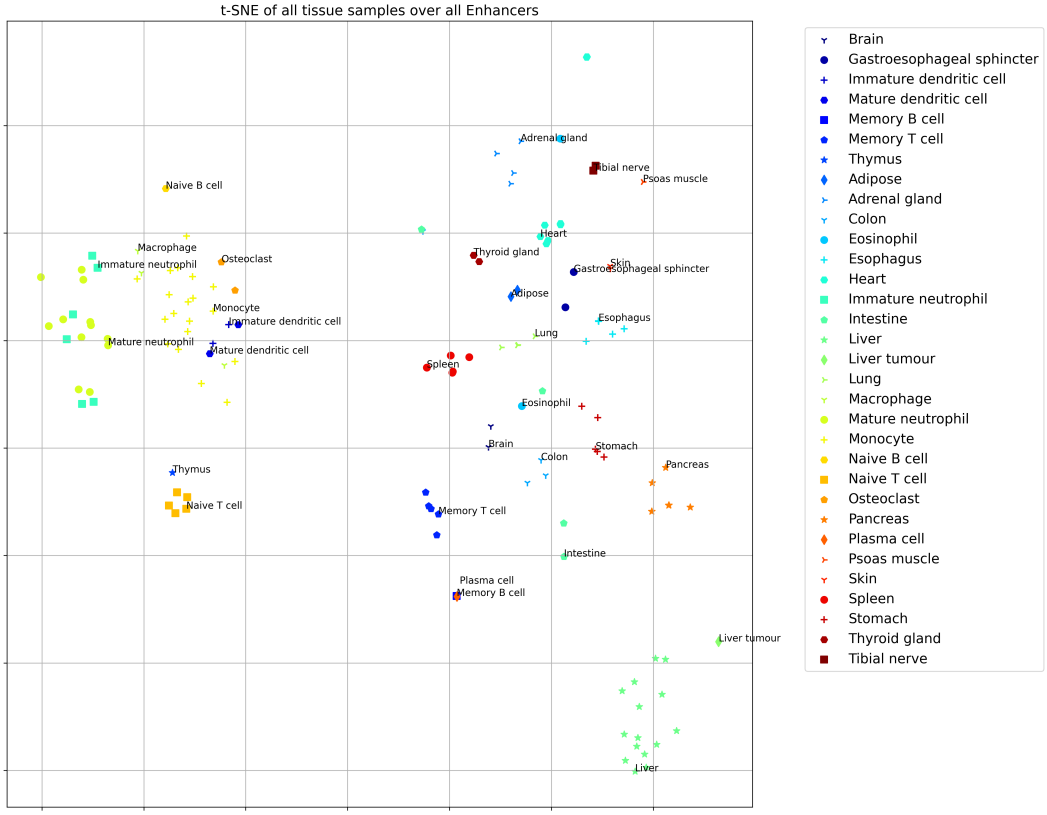


Figure 3.4: All tissue samples in the final reference atlas projected into 2D space using a t-SNE embedding of methylation across all Enhancers.

3.3.2 Identification of tissue-specific differentially methylated regions

A natural first step in analysing this tissue methylation atlas is to search for tissue informative regions, in this case enhancers demonstrating tissue-specific differential methylation. We are interested in filtering enhancers to find the most tissue specific regions, in order to obtain tissue specific signatures that can be used in our reference atlas for tissue deconvolution of cell-free DNA. A number of pub-

lished DMR finding algorithms were evaluated, which are discussed in more detail in the Background chapter. After reviewing and testing various published methods, I found the most suitable and effective DMR-finding method to be a median difference-based algorithm. This method is similar to the DMR-finding algorithm used in the MethAtlas deconvolution paper [85], but with some additions that improve the tissue specificity and biological plausibility of resulting DMRs. The algorithm is summarised as follows:

For each tissue in the reference atlas:

1. Calculate median methylation values across for all tissue samples in that tissue group (the target group), vectorised across all regions (enhancers in this case).

2. Calculate median methylation of all other tissue samples (the background group)

3. Calculate a methylation delta between the two median values, and rank by this delta.

4. Repeat for all tissues in the atlas.

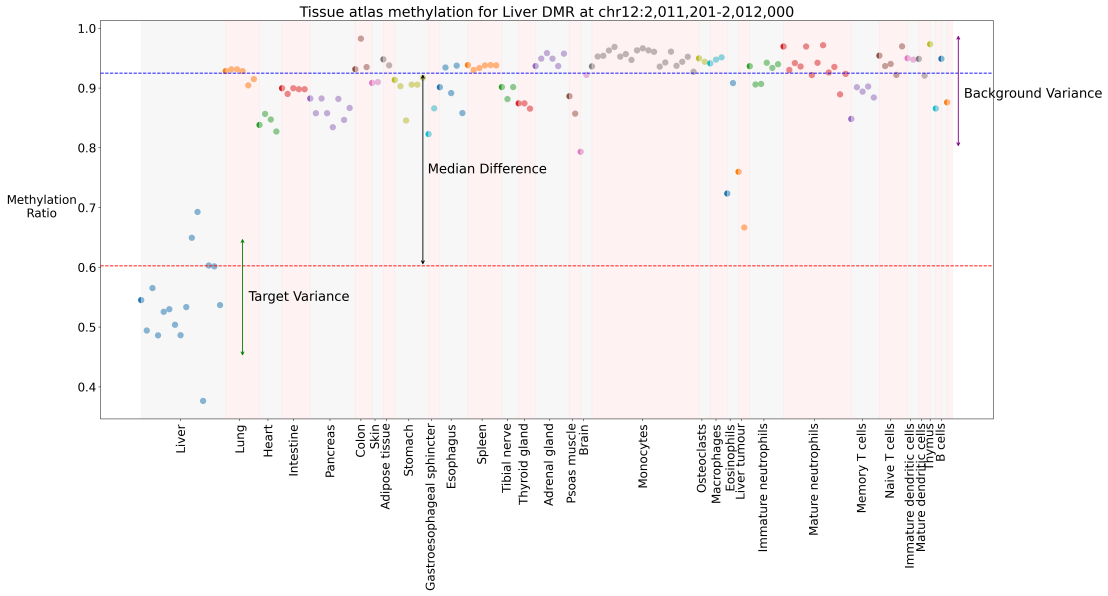


Figure 3.5: The key DMR filtering criteria I used, illustrated at a single genomic locus containing a (real) liver DMR. Arrows show the median difference and the variance of the target group (liver in this case) and background group.

This DMR-finding process is illustrated in the schematic in Figure 3.5. The additional tweaks I made to this algorithm were in the form of optional extra filtering steps, as follows:

1. Set hard thresholds on median values for target or background group, for example for a hypo DMR, requiring median methylation value < 0.1 for the target group and median methylation value > 0.9 for the background group. These thresholds were set according to the tissue methylation distributions.

2. Requiring target or background group to have lower than a certain variance, to only retain DMRs with a consistent background methylation state such as constitutive methylation.

3. For each tissue group in the atlas, identify the two most closely related tissues (in terms of overall methylome signature) and add pairwise DMRs to better distiniguish these closely related tissues.

Filtering steps 1 and 2 can be seen as introducing a biologically informed constraint, by essentially selecting for regions with fairly pure and stable methylation state in the target group, thus enriching for stronger hypomethylation events. We focus on hypomethylated DMRs here, as this was the primary type of DMR detectable in this public data.

3.3.3 Evaluating DMR tissue-specificity

There is no ground truth set of tissue specific DMRs here, so it is challenging to quantify the resulting degree of enrichment in tissue informative markers. It is possible to assess tissue specificity of DMRs by visually inspecting heatmaps of methylation across all tissues, but this method is fallible and requires quantifying. I identified statistics originally devised for economics but which have been used to quantify tissue specificity of gene expression, and adapted these for methylation data. These are Gini impurity coefficient [131] and Shannon entropy specificity [132]. Their adaptation to measure tissue specificity of DMRs is described in the Background chapter, but the intuition behind these scores are provided here.

The Gini coefficient is a measure of inequality across groups. A high Gini coefficient here would therefore indicate that few tissue samples have very distinct methylation levels compared to the rest, thus indicating a more tissue specific DMR. Relatedly, Shannon entropy is a measure of information entropy in a dataset, which in this case is just the variance of methylation levels across a region. In this case, because Shannon entropy specificity is calculated by subtracting calculated entropy (variance) from the number of tissues, a high score indicates a region with less variance and higher tissue specificity.

To further evaluate tissue specificity of DMRs, I took the intersection of each DMR set on the tissue samples, and clustered the intersection using unsupervised hierarchical clustering. I quantified overall quality of clustering using the unsupervised (meaning tissue group labels not provided) Silhouette Score metric [133], then used two supervised metrics to quantify the extent to which tissue samples clustered into their tissue groups. These were adjusted Rand index (ARI) and normalised mutual information (NMI) score [134, 135]. These metrics are also summarised in the Background section. A high Silhouette score indicates a clearer

assignment of tissue samples into groups, whilst high ARI and NMI scores both indicate better agreement of the assigned clustering with the true underlying groups.

To evaluate whether my DMR finding approach enriched for tissue specific DMRs, I compared these five metrics between the original dataset of all 87,000 Enhancer regions, my DMRs, and an implementation of MethAtlas DMR finding algorithm, which was the best available alternative method at the time. Scores are summarised in Table 3.1. The tissue markers identified by DMR finding method show improved tissue specificity across all five scores compared to the alternative regions. Figure 3.6 shows the final set of DMRs used, and Figure 3.7 shows a t-SNE projection of the final reference methylation atlas after filtering by tissue DMRs using this method, in which we can see improved tissue-specific clustering.

	All Enhancers	**MethAtlas DMRs**	**My DMR method**
Gini Impurity	0.086	0.132	0.139
Shannon entropy specificity	0.031	0.064	0.070
Silhouette score	0.214	0.202	0.266
Normalised mutual information	0.795	0.789	0.826
Adjusted Rand index	0.091	0.058	0.358

Table 3.1: Comparison of tissue specificity metrics for different DMR methods. For each metric, a higher score indicates improved tissue specificity.

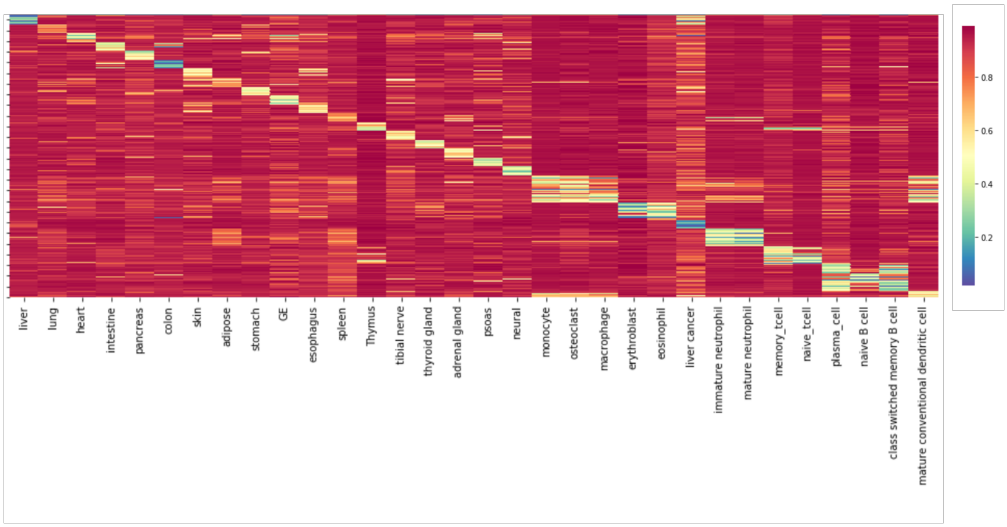


Figure 3.6: A heatmap of the final set of DMRs obtained from the reference atlas, after optimising the DMR calling algorithm to extract maximally tissue informative regions.

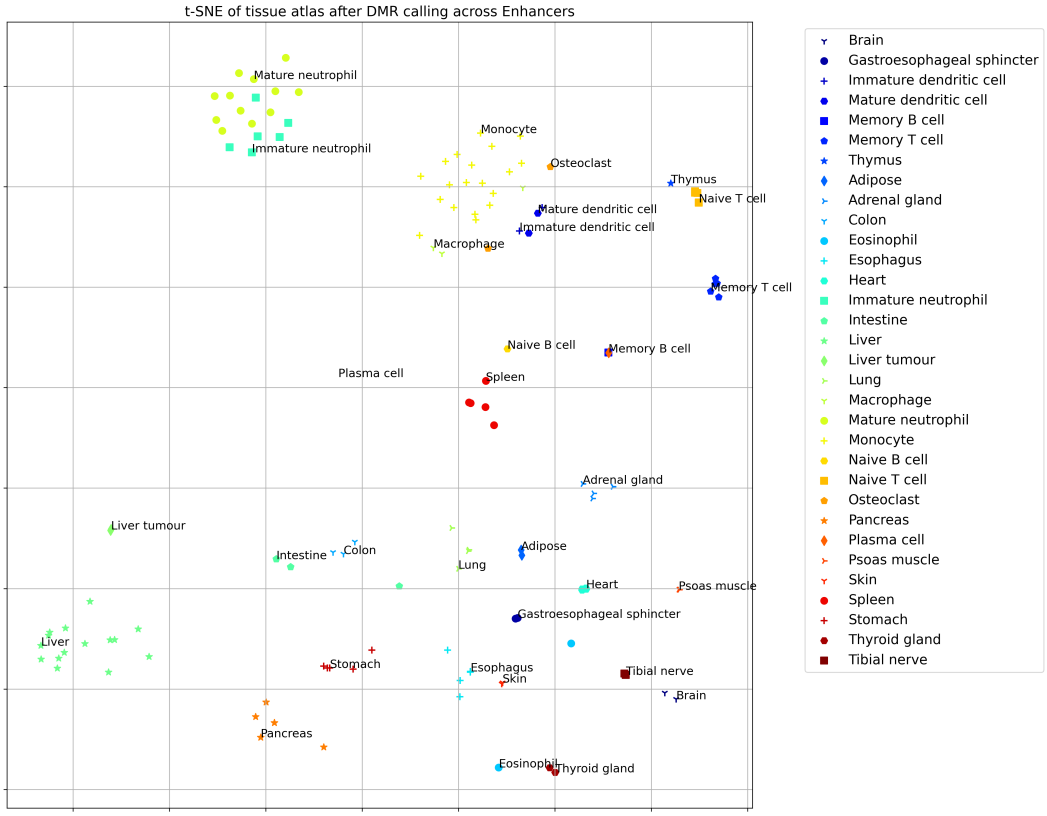


Figure 3.7: A t-SNE projection of the tissue atlas intersected on the final set of enhancer DMRs, which are used for downstream tissue deconvolution.

3.4 Tissue deconvolution of cell-free DNA using a methylation reference atlas

Having assembled this reference atlas and enriched for tissue-informative methylation markers, we are now in a position to use this data as a reference for tissue deconvolution of cell-free DNA.

3.4.1 Cell-free DNA patient cohort

The cell-free DNA of 85 patients was extracted and sequenced as part of a TAPS trial cohort. This cohort included samples from 21 patients with HCC, 23 with PDAC, 30 noncancer controls, 4 patients with cirrhosis, and 7 with pancreatitis, as summarised in Figure 3.8. Cirrhosis and pancreatitis are precancerous conditions affecting the liver and pancreas, respectively. PDAC patients were consented for this study via the Oxford Radcliffe Biobank (09/H0606/5+5, project: 19/A177). Blood samples from patients with clinically diagnosed Pancreatitis, HCC and Cirrhosis patients were recruited for this trial at the John Radcliffe Hospital. The study was approved by Oxfordshire Research Ethics Committee A (REC-A) (10/H0604/51) and is registered on the U.K. National Institute for Health Research (NIHR) portfolio as study number 10776. Notably, the noncancer controls were collected from an endoscopy clinic and were enriched with gastrointestinal inflammatory conditions such as Crohn's disease and colitis. While distinguishing these noncancer controls from cancer patients is more challenging than a typically healthy control group, this may provide a more real-world comparison of a diagnostic test in an aging population. Full patient information is provided in Tables 1-3 in the Appendix.

Most PDAC and HCC patients in our cohort were at a nonmetastatic stage, with 52% of PDAC and 67% HCC patients at stages I and II. Among the 21 HCC patients, only 4 (19%) had elevated levels of AFP (over 20 ng/ml). Among the 18 PDAC patients who had CA19-9 measurement, 16 (89%) had elevated levels of CA19-9 (over 37 U/ml). However, CA19-9 level is often elevated in nonmalignant conditions including inflammatory disease.

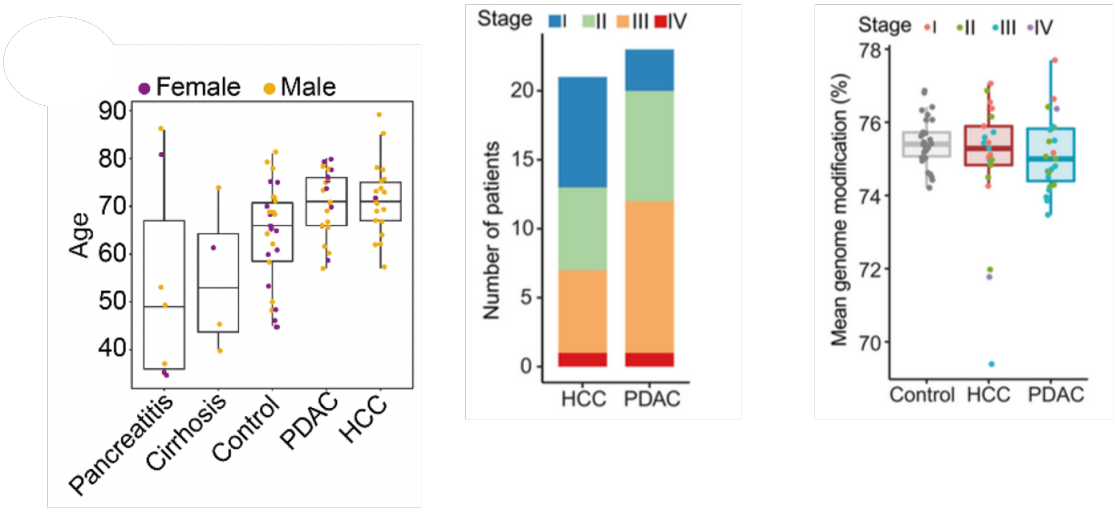


Figure 3.8: Metadata for this cfDNA patient cohort. Left: age and sex distribution in each of the five groups. Middle: cancer stage distribution for the HCC and PDAC cancer patients. Right: mean genome modification for the two cancer types compared to non-cancer controls.

3.4.2 Tissue deconvolution by non-negative least squares regression

I applied NNLS deconvolution to the methylation signature of our 85 cfDNA samples, taken over all tissue specific Enhancer DMRs as described above. The inferred cfDNA tissue contribution was broadly similar between cancer and control groups, in agreement with previous reports [85, 64, 91], with blood and immune cells dominant, and lower proportions of solid tissues (Figure 3.9). The inferred contribution was highly variable across each group, including the non-cancer control. I observed a significantly increased liver tumor contribution in HCC alone (paired t test, P = 0.0016) and a significantly increased memory T cell contribution in PDAC samples (paired t test, P = 0.028). I did not see any significant increase in pancreas contribution within PDAC samples compared to non-cancer controls.

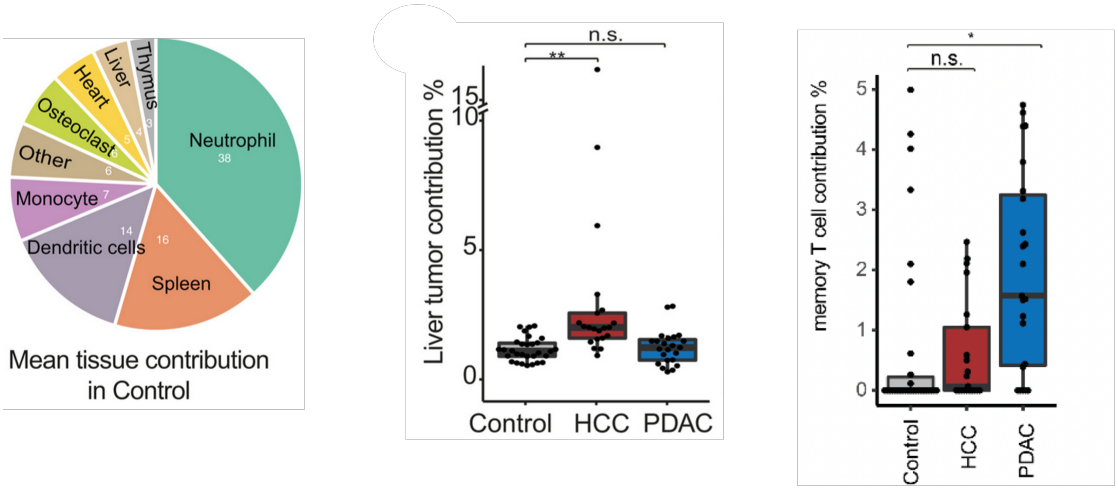


Figure 3.9: Left: mean tissue contribution in control group, dominated by blood and immune cell types. Middle: liver tumour contribution across the three groups, demonstrating a statistically significant increase in liver tumour contribution in HCC. Right: T cell contribution across three groups, demonstrating a statistically significant increase in T cell contribution assigned to PDAC cfDNA but not to HCC cfDNA.

I trained a RandomForest classifier to distinguish healthy cfDNA from cancer based on tissue contribution alone, evaluated by 5-fold cross-validation. For HCC versus noncancer control this classifier achieved a mean AUROC of 0.88 (SD over 5 repeats = 0.14); For PDAC versus noncancer control this classifier achieved a mean AUROC of 0.86 (SD over 5 repeats = 0.15). The misclassified cfDNA samples are from patients with earlier stage cancers, where tumour shedding and physiological changes are less advanced. This is not a thorough evaluation with an independent held-out test dataset, but this does nonetheless demonstrate that cfDNA tissue deconvolution, using this pipeline of enriching for tissue specific DMRs followed by NNLS, provides valuable information that can help distinguish cancer cfDNA samples from non-cancer controls.

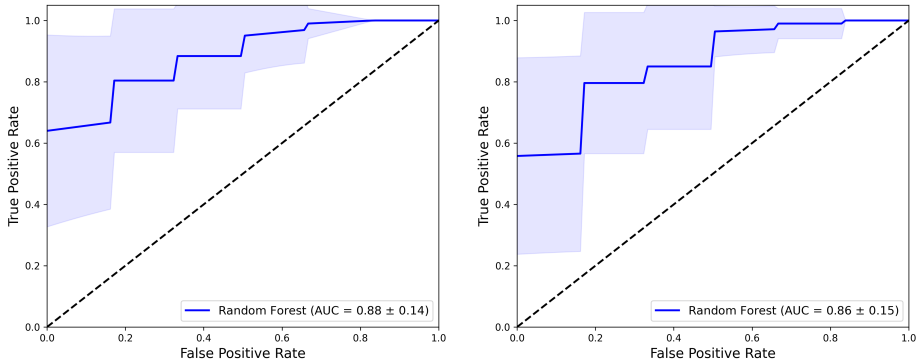


Figure 3.10: Mean tissue contribution across the three groups as inferred by NNLS

3.4.3 Tissue deconvolution of cfDNA using CelFiE

CelFiE[121] is a probabilistic algorithm that uses expectation maximisation to deconvolute tissue contributions in cell-free DNA. I benchmarked different CelFiE model parameters on simulated mixtures of cfDNA, then I repeated the above deconvolution experiment, using the same input datasets. Some theoretical benefits to using this approach were not seen in practice when implementing CelFiE deconvolution. For example, whilst there is no theoretical limit to the number of CpGs considered in the input data, expectation-maximisation models converge more easily and achieve a stable deconvolution result across repeated runs when 5000 odd tissue informative markers are used.

An additional parameter in CelFiE tissue decomposition which the user can specify at runtime is the 'number of unknown tissues' present in the cfDNA sample for which there is no reference tissue. This is a theoretically very useful addition to mitigate reference bias in reference-bsed deconvolution. However, the exact interpretation of this parameter is debatable: should the number of unknowns reflect all possible tissue contributions that are not present in the atlas, or just a single major cfDNA contributor known to be missing from the reference in a specific context? After corresponding with the authors, I varied the number of unknowns between zero and two as is done in the original study. Likelihood estimation does not proceed stably if many unknowns are stipulated in the reference, even though this is the reality for tissue deconvolution with limited reference data. This does

not unfortunately solve the reference bias issue, as the nature of cfDNA deconvolution as a blind source separation problem means we cannot reasonably know we are missing one particular tissue in the reference. The number of unknowns parameter may be more usefully intepreted as a measure of 'missingness' in the reference, ie the number of latent signatures contained in the cfDNA that could not be confidently assigned to a signature in the reference provided.

Including 'unknown' contributions alters the absolute proportions of each tissue contribution estimate, as shown in Figures 3.11 and 3.12. A similar general trend is observed to the above NNLS contribution, where blood cells comprise over 50% of 'known' contributions, however an important difference is evident in estimates of overall tissue contributions between the two algorithms: CelFiE estimates small but non-zero contributions for the majority of tissues in the atlas, whereas NNLS has a tendency to assign weak signals to zero, resulting in fewer, larger tissue contribution estimates.

The tissue contributions estimated using CelFiE are largely similar to those estimated with NNLS. An increased liver, B cell contribution is observed in HCC and PDAC samples, and increased liver cancer contribution is observed in HCC samples. A more consistent increase in PDAC-specific pancreas tissue contribution is observed in CelFiE, as illustrated in Figures 3.11 and 3.12. Comparing tissue contribution assigned by CelFiE versus NNLS, we see that the CelFiE algorithm has a preference to assign small but non-zero contributions for the majority of weak signals, whereas NNLS assigns these small contributions to zero.

I systematically compared all tissue contributions across CelFiE versus NNLS in Healthy (Figure 3.13) and HCC (Figure 3.14), and this trend can be clearly seen. I tested whether the difference between the assigned contributions of the two algorithms were statistically significant, using a Mann-Whitney U test with Bonferroni correction for multiple testing. As suggested by the data distributions in the boxplots, the majority of the differences were significant for both Healthy and HCC, except for monocyte and osteoclast tissues. Overall, these two deconvolution algorithms largely agree in trend of tissue contribution, assigning the majority contribution to blood cells such as neutrophils, dendritic cells and monocytes, although the exact ordering is different.

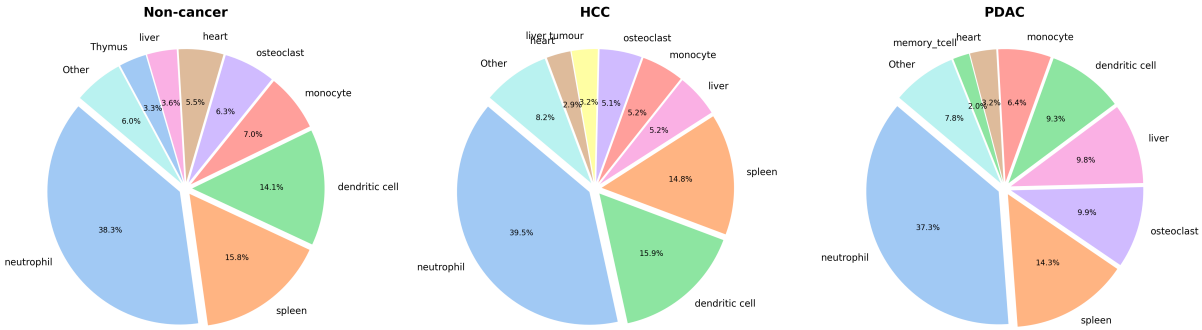


Figure 3.11: Mean tissue contribution across the three groups as inferred by NNLS

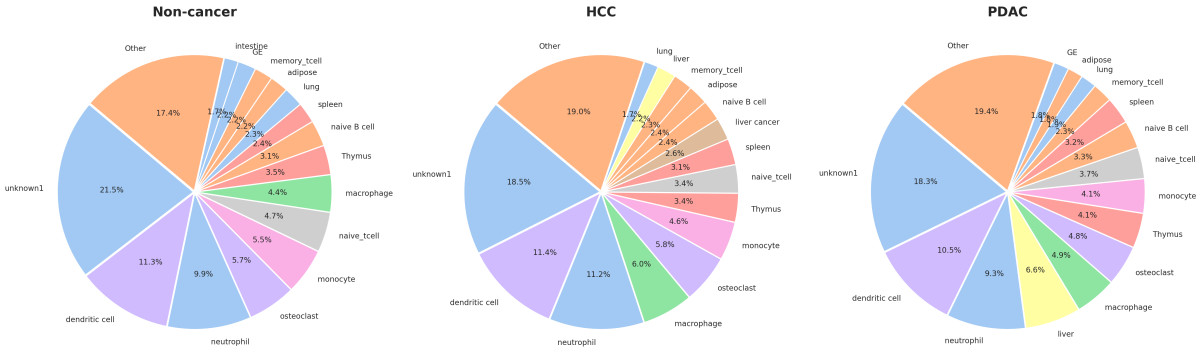


Figure 3.12: Mean tissue contribution across the three groups as inferred by Celfie, with one unknown stipulated.

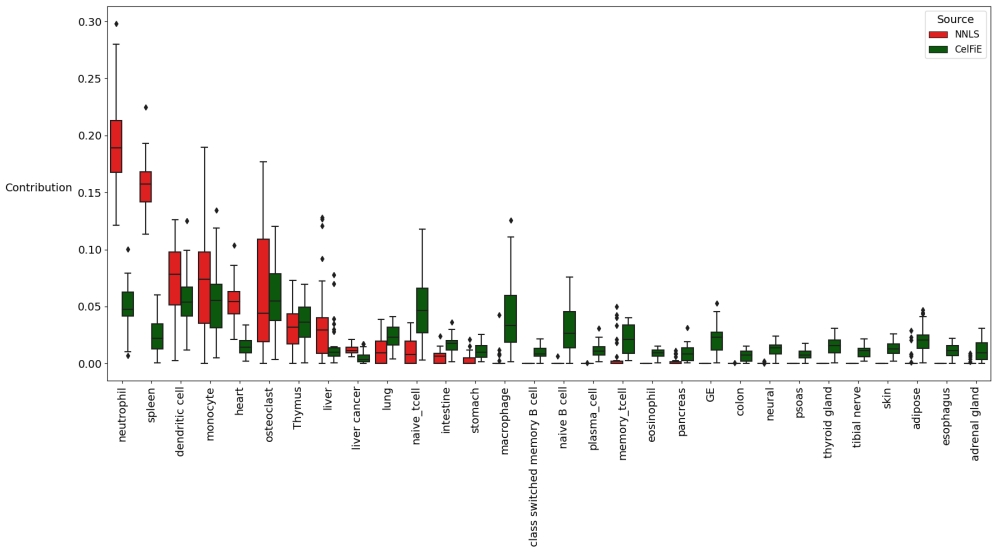


Figure 3.13: Boxplot comparing tissue contribution across healthy samples, as estimated by NNLS and CelFiE algorithms. Boxplots show median and interquartile range, with outliers labelled as dots.

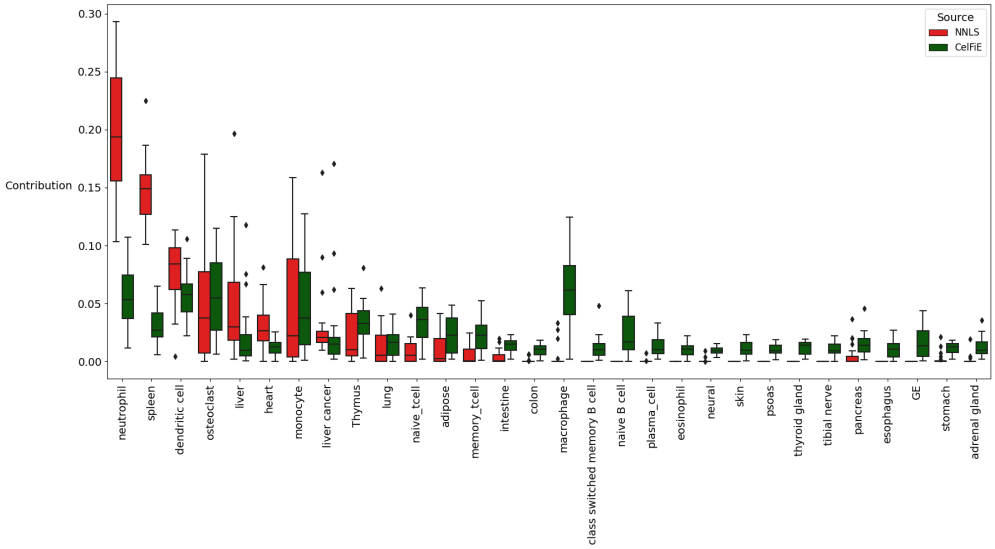


Figure 3.14: Boxplot comparing tissue contribution across HCC samples, as estimated by NNLS and CelFiE algorithms. Difference between model estimates for most tissues were statistically significant, except for osteoclasts and monocyte.

3.5 Developing a deep learning autoencoder for cfDNA tissue deconvolution

To deconvolute cfDNA, I use a deep autoencoder neural network, composed of an encoder and decoder. The encoder learns to take a simulated cfDNA matrix X and to predict a vector of proportions y (the relative tissue contributions in this simulated cfDNA). The decoder takes this predicted vector of tissue contributions, and attempts to recapitulate the simulated data input X. We start with a tissue reference matrix of methylation values (between 0 and 1) of k tissue types across m genomic positions (CpGs). A more comprehensive methylation atlas was published by the time of this work, which demonstrated higher genome coverage and cell-type resolution for many organ systems [91], so I used this reference to evaluate this model. The model architecture used here is inspired by two recent gene expression deconvolution methods [138, 137].

I first simulated the human cfDNA pool by mixing tissue-specific methylation marks from a real biological reference atlas, according to a beta binomial distribution, which is a widely used distribution for modelling tissue methylation across genomic CpGs [116, 151]. We simulate mixing at a vector of pre-defined proportions, which becomes the target variable y that we aim to predict. The simulated cfDNA matrix is the input training data, X.

We train the encoder to take this simulated cfDNA input and learn to correctly predict the known vector of proportions y, by minimising the mean absolute error between predicted and true vector of proportions. As usual for autoencoders, we also incorporate the reconstruction loss into the encoder training. The encoder cost function therefore incorporates two losses: the classical reconstruction loss, and the supervised loss of predicting a known vector of tissue proportions y. We subsequently optimise the decoder to reconstruct the original cfDNA input. A useful property of this architecture is that the weights of a well-optimised decoder correspond to the underlying tissue specific methylation signatures. The model architecture and associated losses are illustrated in Figure 3.15.

I simulate n=80,000 unique mixtures of cfDNA, and train this network across 200 epochs using the Adam optimizer with the default learning rate of 0.001 and batch size of 128. The autoencoder model is symmetric and uses five dense, fully

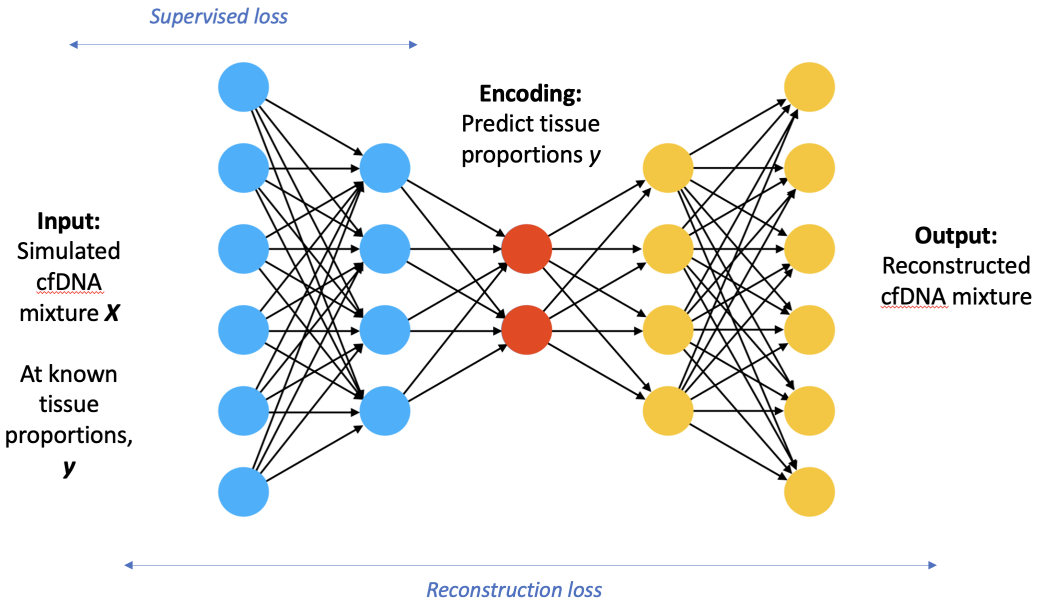


Figure 3.15: Autoencoder schematic, showing input, output and model loss. Also shown are the two separate losses as used in model training.

connected input and output layers, each followed by batch normalisation to improve gradient flow through the network and ReLU activation to learn non-linear mappings. Due to resource constraints, hyperparameter optimisation was not performed here; hyperparameter optimisation in future iterations may improve the model performance further. PyTorch (version 1.12.1) was used for all models.

3.5.1 Training on simulated data

The trained autoencoder successfully deconvolutes simulated cfDNA, by reconstructing the original simulated input cfDNA mixtures X with high accuracy, as well as the assigned tissue proportions y. To measure this reconstruction loss, we need to compare the original cfDNA matrix to the reconstructed matrix. I use mean Spearman's correlation averaged across each sample n throughout.

The tissue contribution to cfDNA is highly variable across both tissues and time, and any deconvolution method must be robust to this variation. I therefore measured the performance of this deconvolution method across three biologically plausible scenarios. First, I simulated a randomly drawn split in contribution

across all tissues, where each tissue has an equal expected value for assigned proportion. I repeated the data simulation and model evaluation 30 times, as in [121], and report the mean correlation coefficient and standard deviation. Our method has a similar mean correlation to NNLS for this scenario, as shown in Table 3.2.

Scenario	**NNLS**	**AutoEncoder**
30 repeat simulations	0.941 (SD=0.003)	0.965 (SD=0.008)
Variable contribution	0.74 (SD=0.005)	0.78 (SD=0.01)
Tissue injury	0.69 (SD=0.004)	0.83 (SD=0.009)

Table 3.2: Spearman correlation with simulated test data for the two methods across a range of scenarios

The second scenario that I simulated is a patient with highly variable tissue contribution, where ten tissues do not contribute detectable levels of cfDNA. I encoded robustness into the autoencoder deconvolution model by training with a wide range of tissue contribution scenarios. The subsequent trained model outperformed NNLS when comparing to the true underlying tissue proportions in the simulated test data.

Finally, I simulated the scenario of patient tissue transplant. This physiologically resembles acute tissue injury, where cfDNA from the newly transplanted organ is abnormally high in the recipients blood, and which subsequently varies according to the degree of transplant rejection [64]. I simulated this by assigning 60% of total cfDNA contribution to a single tissue, and trace amounts to the remaining tissues. The autoencoder model outperformed NNLS here, as summarised in 3.2. It should be noted that this method also showed a higher average variance across repeated simulations compared to NNLS, as indicated in the higher standard deviations of the observed correlation.

3.5.2 Deconvolution of patient cell-free DNA

After demonstrating that the model was able to accurately deconvolute simulated cfDNA, I sought to apply this trained model on the 51 HCC and non-cancer con-

trol cfDNA samples used previously. The tissue and cell-type contribution inferred by the trained autoencoder model agree well with both the previous deconvolution results, and the general literature consensus [121, 64, 85]. The major contributors to the inferred cfDNA pool are proximal blood and immune cell types, such as neutrophils, monocytes, megakaryocytes and immune lymphocytes, as shown in Figure 3.16. I similarly trained an autoencoder deconvolution model on the previous WGBS reference I collated, and obtained a comparable result.

As discussed, cell-free DNA shedding is highly variable and dependent upon patient-specific physiology, so the high individual variability in tissue contribution within both groups is expected, especially as the non-cancer controls were not a typical healthy cohort. Nonetheless, important differences between control and liver cancer groups can be discerned, as shown in Figures 3.16 and 3.17. Firstly, an overall increased hepatocyte contribution to cfDNA is inferred for liver cancer patients compared to non-cancer controls. Secondly a slighty increased T cell contribution is also observed in HCC samples. The inferred tissue contribution differences are sufficient to distinguish HCC from control using an out-the-box implementation of a RandomForest classifier from scikit-learn with similar accuracy as previously (Random forest AUC: 0.94 with std dev. 0.11 over 10 repeated cross-validation folds).

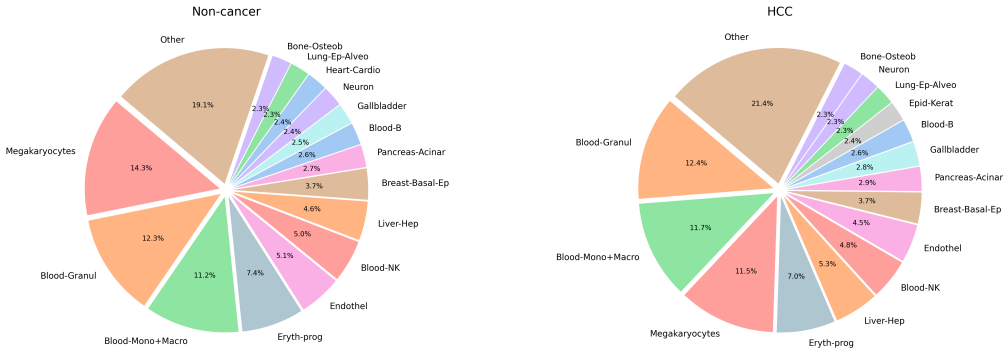


Figure 3.16: Cell type contribution averaged across each of the two groups, liver cancer (HCC) and non-cancer controls.

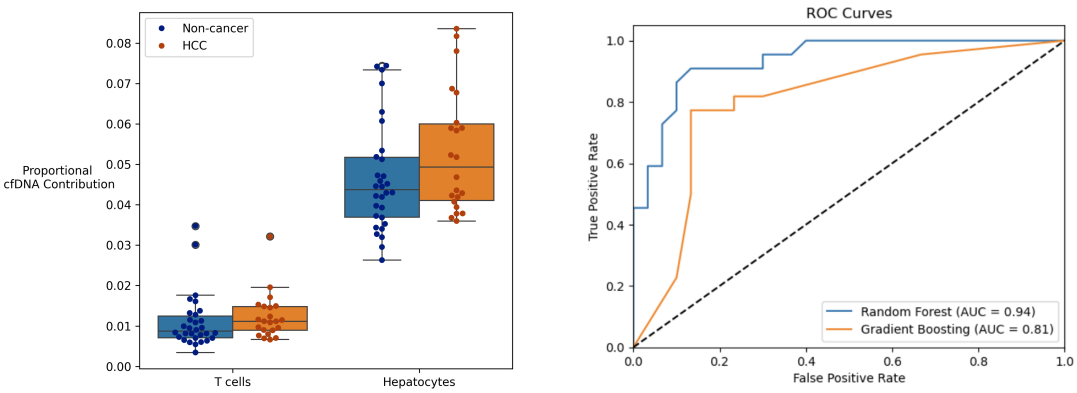


Figure 3.17: Left: Differences in inferred T cell and hepatocyte contribution between HCC and non-cancer cfDNA samples. Right: ROC curve for random forest and gradient boosting classifiers trained to distinguish HCC from non-cancer samples using the inferred contributions. Random forest AUC: 0.94 with std. dev 0.11; gradient boosting AUC: 0.81 with std dev 0.09

3.5.3 Reconstruction of cell-type specific methylation signatures

I next sought to reconstruct the original methylation signatures in the reference atlas. In learning to reconstruct the cfDNA input, the weights of the decoder can be taken as cell-type specific methylation signatures. When comparing these inferred signatures to the true tissue methylation reference data, I found that each inferred signature generally correlated best with the correct original tissue reference profile, as shown in Figure 3.18.

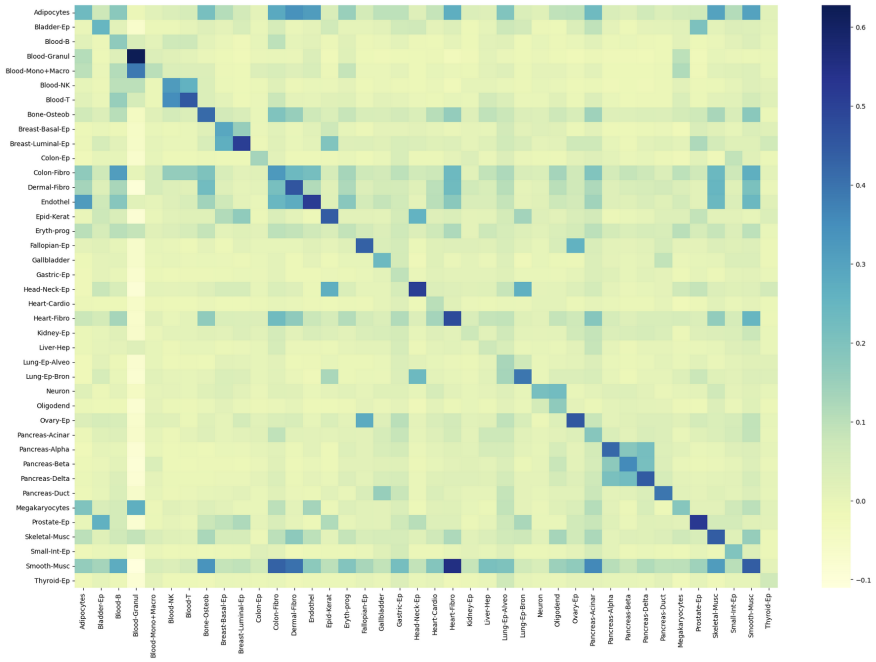


Figure 3.18: Pearson correlation of reconstructed cell-type specific methylation signatures with true reference signature.

These are the reconstruction results after a single round of decoder optimisation as a proof of concept. Multiple rounds of iterative optimisation of the decoder and the encoder improve the reconstruction loss, and thus improve the reconstruction of these tissue-specific methylation signatures.

3.6 Discussion

This chapter focuses on optimising all the steps in the pipeline for reference-based deconvolution of cell-free DNA. To minimise bias introduced by the reference, I maximised both the number of tissues captured in the reference, and the breadth and depth of genomic sequencing coverage for each tissue. By collecting multiple biological replicates per tissue and removing anomalous samples that exhibited batch effects or otherwise clustered separately from tissue centroids, we increase the likelihood of capturing a truly representative methylation signature for each tissue. At the time of this work, this was the most comprehensive genome-wide tissue methylation reference. However, several larger methylation references have since been built, most notably the cell-type atlas released by Loyfer *et al* in 2023, which I use for benchmarking the autoencoder based deconvolution approach. This demonstrates the current interest in assembling large tissue references for methylation, and their importance for cfDNA deconvolution.

All tissue deconvolution methods currently in use still require a manual feature selection step to enrich for tissue informative markers before deconvolution. The DMR calling method I develop and use here seems to enrich for tissue-informative markers, and the introduction of additional biologically informed constraints does seem to improve the tissue specificity of the resulting reference, based on improved clustering and specificity metrics. There are many more sophisticated DMR calling algorithms developed for pair-wise case versus control problems, but the stable and consistent results we see here validate this feature enrichment pipeline for downstream deconvolution.

Future deconvolution algorithms may at some stage remove the need for this manual feature selection step of narrowing down the regions of the genome considered before performing deconvolution. There is no theoretical limit to the number of genomic features that could be considered in such a model; the guiding principle here is the empirical fall-off in model accuracy scores as the signal to noise ratio decreases. Deep learning methods excel at automated feature selection: by backpropagating model coefficients onto important features and ignoring irrelevant features, they have removed the need for manual feature selection in various

domains. Now we have a working proof of concept for deep learning based deconvolution, it is possible that future iterations replace this manual feature selection step with end-to-end fully connected networks and gradient-descent.

I performed cfDNA deconvolution with three distinct methods and two separate references, and it is noteworthy that the inferred tissue contributions largely agree across methodologies. Blood and immune cells consistently dominate tissue contributions, as is seen in the broader literature, and solid tissues make up a relatively small proportion of total contribution. We consistently infer a statistically significant increase in liver-tumour contribution in HCC samples, which is useful for distinguishing cancer patients from non-cancer controls. When liver-tumour was missing from the reference, we instead observed a HCC-specific increase in hepatocytes. Interestingly, we did not observe a similar increase in pancreas contribution in PDAC samples. We cannot confidently say whether pancreas specific cfDNA is not present in cfDNA or is simply not detectable. However, this could be due to the fact that pancreatic ductal adenocarcinomas do not shed cfDNA as readily as HCC. A subset of late-stage PDAC cfDNA samples also exhibited high liver contribution. Personal correspondence with clinicians involved in this pilot suggested this may be due to genuine liver inflammation or even liver metastasis in late stage pancreatic cancer, but more experiments would be needed to untangle this.

We also observe fairly large within-group variability across all deconvolution experiments. Cell-free DNA is a highly dynamic biomolecule, and the tissue-specific rates of shedding as denoted in frequency matrix F are in a constant state of flux depending on individual tissue physiology [152]. The temporal fluctuations in cfDNA should also not be under-estimated here. Patient cfDNA samples were collected at a single time point, so necessarily represent a single snapshot of what is in fact a distribution. Physiological differences between patients at this time point, such as time since last exercise, diet and inflammation can have large effects on cfDNA shedding [153]. Collecting longitudinal samples through time would allow us to capture the patient-specific distribution of this naturally fluctuating biomarker, and establish a patient-specific baseline against which to identify malignant changes.

It should be noted that individuals in the non-cancer arm of this pilot study were not a typical 'healthy' cohort. Many had severe inflammation based comorbidities, which may partially explain why several non-cancer cfDNA samples were consistently assigned high T and B lymphocyte contributions. We observed increases in T and B cell contributions in HCC and PDAC cancer groups, but no statistically significant increases that were consistent across this fairly small sample size. Nevertheless, the distinct immune cell contribution profiles in these cancer samples allowed us to accurately distinguish HCC and PDAC from non-cancer controls when using classifiers trained on the entire contribution profile.

These increases in immune cell contributions would perhaps be more pronounced in a study with a more typical control arm. This would be interesting to follow up in a larger study with more patients, as dramatic changes occur in immune cell activity occur during tumourigenesis, related to inflammation. Furthermore, specific immune cell dynamics are associated with specific cancer aetiologies, such as an increase in neutrophil to lymphocyte ratio in hepatocellular carcinoma [154]. A unique advantage of genome-wide cell-free DNA tissue deconvolution compared to targeted assays is this assay is fundamentally exploratory. It enables us to take a more holistic view of the overall physiological changes occurring in cancer patients, rather than focusing purely on detecting tumour DNA. Recognising distinct and cancer-specific immune cell changes are likely to prove an important part of this puzzle.

The performance of the deep autoencoder model demonstrates the potential of re-framing this challenging computational objective as a supervised learning problem. Any deconvolution method must be robust to a range of biologically plausible scenarios. By incorporating supervised learning, we can train model robustness by simulating a wide range of biologically plausible scenarios in the model training data, and evaluate the model performance in each scenario. This evaluation effort is ongoing, and I expect to find further advantages and pitfalls for this deconvolution approach as I scale up this model. It should be noted that a month after this paper was submitted, a new method using supervised deep learning for cfDNA deconvolution was published [155]. At the time of submission, this was still a novel, untested idea. The authors demonstrate the efficacy of a very similar method, thus validating this general approach.

A persistent problem of applying machine learning in clinical contexts is the expense and scarcity of patient data. This scarcity is a major cause of the lack of actually useful machine learning models being applied in healthcare today. In this supervised learning approach, we sidestep this data availability problem by model training on biologically informed simulated mixtures of cfDNA, whilst comparing to true references to ensure that the model remains biologically meaningful. It is noteworthy and perhaps surprising that a deep learning model trained entirely on simulated data can be clinically useful in distinguishing cancer cfDNA samples. Crucially, this also allows us to scale up the model by simply increasing the size of simulated training data, and if needed model complexity could increase accordingly. After sufficient scaling and model calibration, this approach may be a useful tool to detect cancer early from a non-invasive blood draw, whilst minimising the use of precious patient data in model training.

A crucial element lacking from currently available tissue deconvolution methods, including the one I developed here, is a robust method of quantifying uncertainty in the assigned contribution. Tissues with assigned contribution lower than 2% exhibited fairly high variance in their assigned proportion, relative to their absolute value. More informative than a single number would be the assignment of an interval set within which they are confident the true contribution lies. Conformal prediction represents a promising method for incorporating such intervals into neural network models [156].

Inference of unique methylation signatures in the decoder enables a form of model calibration: if the inferred methylation signatures were not similar to the original reference, this would indicate a poor model calibration and the need for further model tuning. The current consensus method (NNLS) is currently lacking such a calibration step, which is especially important for this problem of cfDNA tissue deconvolution on real patient data, where we can never know the true underlying cfDNA composition of a patient. We propose using the correlation score of the reconstructed methylome profiles to true tissue methylation signatures as a novel means of calibrating the inferred cfDNA tissue deconvolution of patient data.

In many other applications, introduction of supervised learning and deep neural networks has led to profound improvements in model accuracy and therefore utility.

This is true but less pronounced in healthcare, and should not be regarded as a panacea for this difficult problem. As in other healthcare contexts, the more fundamental limiting factor is availability of high quality data, and generalisability across datasets. Nonetheless, re-framing this challenging computational objective as a supervised learning problem does bring along several advantages, such as the ability to scale up model training, and an active field of research dedicated to better uncertainty quantification in neural networks.

Any conclusions on cfDNA deconvolution for cancer detection are fundamentally limited by the fairly small sample size of this pilot study. We observe some promising results, but the necessary next step is to collect more cfDNA data across a wider range of patients to validate these results and examine their generalisability. A particular bottleneck for cfDNA deconvolution is the difficulty of obtaining ground-truth measurements: with current sequencing technology we cannot definitively know the true tissue contributions of cfDNA in an individual. We are limited to benchmarking deconvolution algorithms on simulated data, where it is difficult to obtain a representative comparison with any guarantee that an observed advantage will transfer into real patient data. One possible solution here could be experimental mixing of DNA fragments *in vitro*, over a range of physiologically representative concentrations. These mixtures could act as a type of benchmark with which to compare deconvolution algorithms, the cfDNA deconvolution equivalent of ImageNet for image classifiers [157] or CASP for protein folding [158]. In any case, non-invasive liquid biopsy methods are becoming more prominent in diagnosing cancer and other diseases, and there is a clear need for further evaluation and improvement of deconvolution algorithms to better profile this promising biomarker circulating in the blood.

Chapter 4

Multimodal classifiers of cell-free DNA

4.1 Abstract

Cell-free DNA is a biomolecule containing a wealth of information about the health status, metabolic activity and gene expression program in its tissues of origin. In the previous chapter I focused on epigenetic deconvolution methods for cfDNA, but many other model types are being developed to capture information from cfDNA that can inform patient cancer status, most prominently for liquid biopsy diagnostic tests. Multiple distinct modalities can be measured for cfDNA, including methylation status, fragment length and genomic sequence. However due to technical limitations, this information is rarely captured in a single assay. As a result, few methods exist to combine these distinct modalities into a single model.

TAPS sequencing of cfDNA allows us to capture all this information in a single assay. In this chapter I build multimodal cell-free DNA classifiers combining these data types. I demonstrate that by combining this information we can improve class distinction in a multi-cancer classifier. I then use a novel encoding of this data to develop a fragment-level classifier, which learns to predict probability of cancer origin for individual cell-free DNA fragments. By aggregating these predictions across all fragments, this model can accurately distinguish cancer patients from non-cancer controls across multiple datasets.

4.2 Introduction

Cell-free DNA molecules carry three distinct information modalities, all of which can vary with cancer status: genomic sequence, a methylation pattern and the fragment length, as illustrated in Figure 4.1. A fourth informative datapoint that can be inferred from each fragment sequence is the genomic locus it originated from. There is intense interest in capturing the information contained in this macromolecule, and using it for non-invasive cancer diagnosis; a 2022 review identified 1,370 ongoing clinical trials using cfDNA liquid biopsy for cancer detection in the United States and Europe [76]. However, relatively few investigations have attempted to combine these distinct data types in a single diagnostic model. A key limitation here is technical, where the cost of sequencing necessitates a trade-off between broad and shallow sequencing or deep, targeted sequencing, focusing on a single modality. The harsh reaction conditions of bisulfite sequencing degrade cfDNA, leading to information loss which exarcebates this problem [92].

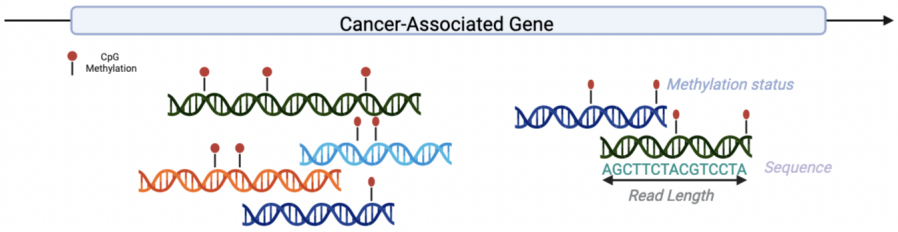


Figure 4.1: A schematic of cell-free DNA fragments found in peripheral blood, aligned to a gene of interest. These fragments contain three distinct modalities of information: DNA sequence, methylation pattern (shown here on CpGs) and their variable fragment length. The genomic locus that each fragment originates from is also informative.

Recently, TAPS and other methods have been developed to simultaneously capture methylation, sequence and fragment length information of cfDNA with high fidelity and depth of coverage, resulting in a rich dataset describing the genome-wide cell-free DNA state of cancer patients [93]. Furthermore, TAPS does not degrade DNA, so maximises the information conversion when sequencing this precious biomolecule. It is therefore possible to build multimodal classifiers of cfDNA

using the data output from this assay. In this chapter I explore the potential utility of such multimodal classifiers. I first build a three-class classifier of cfDNA, combining methylation, fragment-length information and the inferred tissue contribution of cfDNA. This classifier has the more challenging objective of separating control cfDNA from cancer cfDNA whilst simultaneously distinguishing cfDNA from the two cancer types. Incorporating information from these diverse data types bolsters this classifier, which demonstrates improved three-class accuracy.

Patient cfDNA data is scarce and expensive to obtain, so tends to be sample-size limited, as is the case here for our TAPS trial cohort (n=85). We must therefore be careful to avoid overfitting to this small sample size: building complex models that perform well on this data but fail to generalise to new populations. In the second half of this chapter, I focus on developing a fragment-level classifier of cell-free DNA. This represents a shift of focus, where each data point becomes an individual sequenced cfDNA read, rather than the previous paradigm where each data point was an individual patient. This shift in focus allows us to build more complex models that can extract information from individual cfDNA fragments, without overfitting to patient data. Furthermore, this cell-free DNA classifier does not use cfDNA data from cancer patients for model training; models are trained on an adjacent task using public tumour data, and trained models are applied to cfDNA classification by transfer learning.

The first published fragment-level classifier of cfDNA was CancerDetector [124]. This elegant study models the methylation status across all CpGs in each read with a beta-binomial model and uses probabilistic inference to assign reads to one of two classes: healthy origin or cancer origin. They devise a methylation distribution for each of these two classes, healthy and cancer, then use the observed data to calculate the likelihood of the observed joint methylation state across each read, for each class distribution. Taking inspiration from this method, my fragment-level classifier approach also seeks to determine the probability of cancer origin for each read. However, I sought to extend this fragment classifier to incorporate the multiple modalities mentioned above. Other methods have used one-hot encoding to encode DNA sequence as input to a deep learning model, such as DanQ [103]. I extend encoding to incorporate cytosine methylation as a 'fifth base'. A similar approach is taken by DISMIR [101]. By combining multiple modalities in a single

fragment-level classifier, we can start to examine the interplay of these modalities in determining the probability of cancer origin in cell-free DNA.

4.3 Classifying cell-free DNA with a multimodal classifier

Note: the first part of this work on cfDNA fragmentation profiles was a collaboration with J.Cheng and led by P.Siejka-Zielinska.

I sought to develop a classifier that makes full use of the multimodal information contained in cell-free DNA, which is made possible by the technical advance of TAPS sequencing. First, observing that TAPS preserves fragment size information of cell-free DNA, we developed an approach to compare fragmentation profiles of the different groups in our patient cohort. This involves dividing the cfDNA fragmentation distribution into 10-bp bins and calculating the proportion of fragments in each 10-bp bin. The proportion of long fragments (300 to 500 bp) in the overall cfDNA fragmentation profile could be used to distinguish HCC and PDAC from noncancer controls with high accuracy (HCC AUC = 0.92; PDAC AUC = 0.84).

I next investigated various ways of combining a patient's cfDNA fragmentation profile with methylation state and the inferred tissue contribution. This is a challenge of multimodal data integration in a small data regime. I built a multimodal classifier that combined differential methylation, tissue contribution, and fragment profile into a three-way classifier of cfDNA. I use a Support Vector Machine (SVM) classifier implemented in scikit-learn with default parameters and a radial basis function kernel, evaluated using leave-one-out cross validation. The model classifies cfDNA samples as either HCC, PDAC or non-cancer, as summarised in Figure 4.2. This integrated model took the averaged scores from classifiers trained on each of the three modalities and used the most confident prediction for each sample, thus acting as an ensemble model. The overall accuracy of the combined model was 0.86 (64 of 74 were classified correctly), and the accuracy for distinguishing controls from any cancer type is 0.92, which highlights the benefits of incorporating multimodal information for cancer type prediction. The presence of more than one cancer type in this 3-class classification makes it a more challenging task,

and combining multimodal information led to better performance than using any single modality alone.

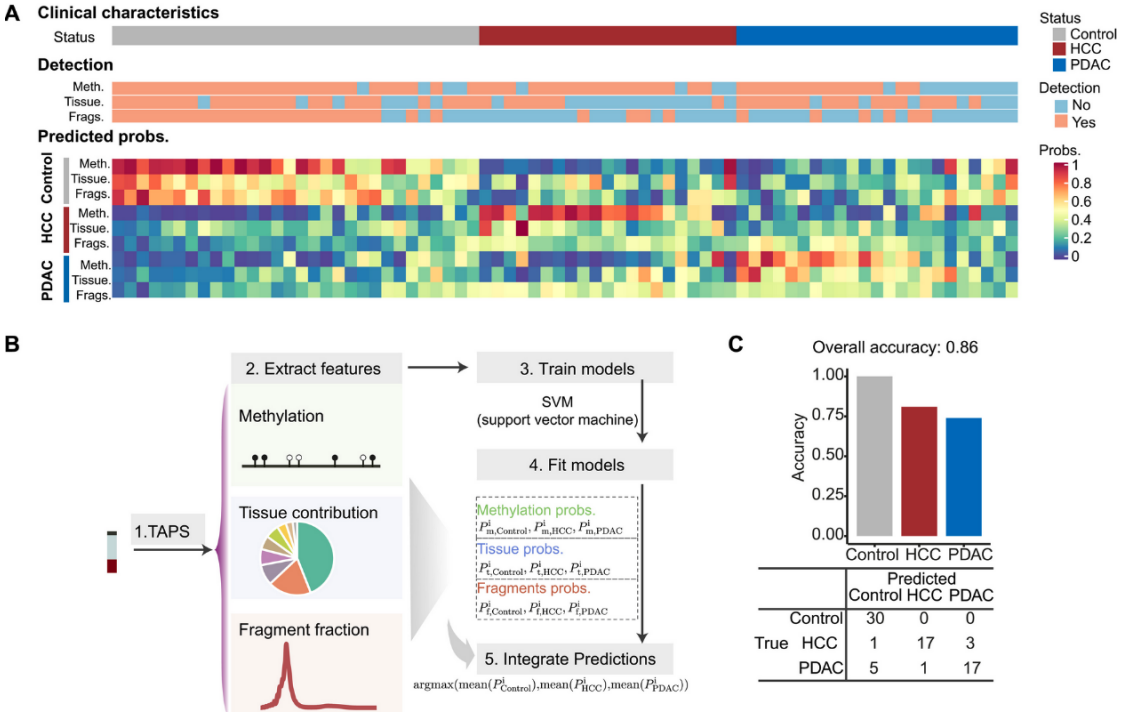


Figure 4.2: (A) Heatmap showing individual model performance on multicancer prediction and the predicted probabilities for each patient. Each vertical column is a patient. Detection yes/no refers to whether patients are correctly classified or misclassified on the basis of a particular feature. Predicted score means the probability of classifying the patients to a specific group based on a particular feature. (B) Schematic detailing the method of integrating multiple features (DNA methylation, tissue contribution, and cfDNA fragmentation) extracted from cell-free TAPS data for multicancer prediction. (C) Overall accuracy and confusion matrix showing the actual and predicted patient status for the ensemble model

We obtain cfDNA sequence data each patient, so in theory can obtain mutations and profile genetic variability across the cohort. However, at the time of this work there remained some difficulties in confidently calling C to T mutations using the TAPS sequencing method, and to avoid the risk of introducing bias, I elected to avoid incorporating mutations. Due to low copy number in early stage disease, mutations are often challenging to detect in cfDNA and require high sequencing coverage. However, multi-class cfDNA classifiers such as this

would likely be strengthened by incorporating genetic information as an additional modality, and this will be a key focus of future work.

This approach demonstrates the utility of combining multiple modalities in cfDNA, but simply combines predictions of each model after classification, so is not a truly multi-modal model. To construct such a model requires us to move away from relational data formatting to a bespoke formatting of the input data to encode spatial components, which itself requires the flexibility of a neural network model for information transfer.

4.4 Developing a read classifier of cell-free DNA

The biological unit of interest in cfDNA is the individual fragment, derived from the nucleus of a bodily tissue and circulating freely in the blood. Each fragment is information rich, as illustrated in Figure 4.1. I therefore sought to develop a classifier focused at the individual fragment level. The model takes individual sequencing reads from healthy cfDNA and tumour samples as input, and learns to separate the two. The trained model is then applied to a distinct task of classifying all fragments in cancer and control cfDNA samples. A probability distribution is thus assigned to all fragments within a cfDNA sample, which represents the probability of cancer read origin for each individual fragment. This distribution is then summarised using maximum posterior probability (as in CancerDetector [124]), to give a patient level classification, as illustrated in Figure 4.3.

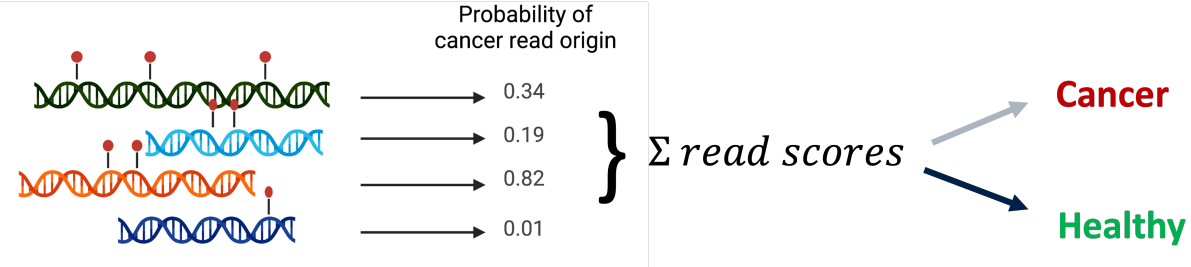


Figure 4.3: Schematic for assigning model scores to individual fragments, which are then aggregated into a patient level classification.

A key challenge is to devise a suitable encoding of information captured on each individual read, such that efficient information transfer could occur through

the model. I encoded each individual cfDNA molecule as a variable length vector, where numbers 1-4 represent each of the possible nucleotide bases at each position. Methylated cytosine is encoded as a 5, to distinguish it from unmethylated cytosine.

To obtain read-level information for each sample, we must start with aligned reads in the (binary) SAM data format. This is a text-rich format that captures comprehensive information from sequencing each read. A key challenge was converting this text data format into a standard relational data matrix where rows represent data points and columns represent features, as needed for model input. I wrote a custom script that parsed each line of this data type, converting each sequenced base to the numerical encoding above. I added custom logic to ensure this read conversion was strand aware, and properly accounted for genomic insertions and deletions. Aligned SAM files of each sequencing read occur upstream of methylation calls, so we lack a labelled value for methylation. I therefore also incorporated logic to identify differentiate methylated cytosine from unmethylated cytosine based on sequence context.

This numerical encoding is converted to a one-hot encoding before passing as data input to the model. At model input, each fragment is therefore represented as a k x 5 array, where k is a fragment's length and the columns represent the one-hot encoding of each of the five possible states at each individual position on the read. In practice, to avoid variable length input, I then added zero-padding to the ends of each array to ensure a standard array length, a common practice for image classifiers.

These encoded molecules are then passed to a deep learning model with a binary classification final layer, whose task is to predict fragment origin: Healthy or Cancer. This model was trained on public data, which consisted of billions of individual reads from either Healthy cfDNA samples or hepatocellular carcinoma tumour samples. To enrich for tumour signal, I took the subset of reads overlapping differentially methylated regions between HCC tumour and healthy cfDNA, using my previous DMR finding algorithm from Chapter 3. I experimented with different neural network architectures best suited to capturing this encoding of fragment-level information, including ConvNets and Transformer models, evaluating performance by nested cross-validation on a subset of samples. Ultimately

the best performing model combined two 1D convolutional layers separated by a bidirectional LSTM layer. The output of each convolutional layer was down-sampled using max pooling, and all layer outputs were passed through a RelU activation function to increase non-linearity of the learnt mapping. Finally, two fully-connected dense layers and a sigmoid activation function were used to reduce the feature representation down to a single output neuron, which represents the calibrated probability of originating from cancer for that fragment. This network architecture performed best in cross-validation and was selected for full model training and evaluation on the transfer learning task on patient cfDNA.

This approach relies on availability of high-quality aligned sequencing data from healthy cfDNA and tumour samples for model training. Public data in this aligned read SAM format is rare, as typically processed outputs of sequencing is shared whilst raw data is protected. Furthermore, publicly sharing this data is challenging because each the file size for each sample is around 150GB. I obtained aligned read .bam files from personal correspondence with Jasmine Zhou and Dennis Lo research groups [82, 124]. I obtained 30 healthy and 24 HCC cfDNA samples and 13 HCC tumour samples for model training and evaluation. Data quality was fairly low, as it was obtained from low coverage WGBS first published in 2013. Importantly, sequencing read-length was 80bp, which is much shorter than read lengths from more recent paired-end sequencing data (average length 150bp). To avoid fragment length discrepancies between data types biasing the model, I limited the model receptive field to the average read length from this dataset.

A further limitation was all data available for model training was WGBS, whereas the cfDNA samples from our pilot study were sequenced using TAPS. I used tumour bam files and healthy cfDNA from this dataset for model training, and evaluated the model in three ways.

1. The held-out test set of unseen cfDNA samples not used for model training, but taken from the same cohort as the training dataset. This consisted of 14 healthy samples and 11 HCC samples.

2. A separate publicly available dataset of cfDNA samples from HCC and healthy controls (HCC n=4, healthy n=4), obtained from the research group behind the 2018 CancerDetector study [124].

3. Our cfDNA samples from the TAPS pilot, consisting of cfDNA samples from HCC and non-cancer controls as previously described (n=53).

4.5 Evaluating cfDNA read classifier on patient data

For any given cancer patient or healthy control, each individual cfDNA fragment is assigned a probability of originating from cancer, as shown in Figure 4.4. As shown, the probability distribution over all fragments is largely similar between healthy and cancer cfDNA samples. This is to be expected: cell-free DNA is thought to be released from most healthy somatic tissues, with the majority contribution coming from healthy lymphocytes. In cancer patients, only a small subset of cfDNA fragments actually originate from the tumour, and even this circulating tumour DNA fraction varies considerably with cancer stage and cancer type. A small bump in the probability distribution around P=0.9 is seen in cancer cfDNA samples (red) but not in healthy cfDNA samples (green), and can be attributed to cfDNA originating from the tumour.

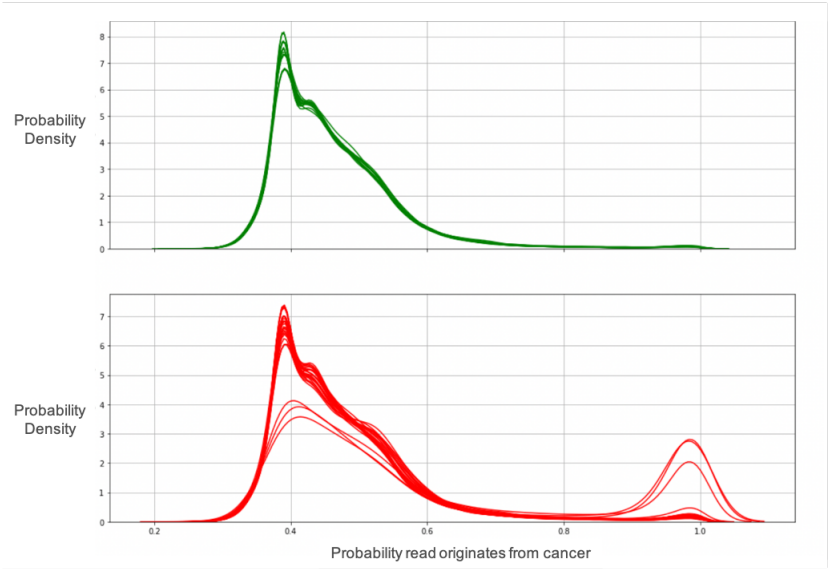


Figure 4.4: The distribution of cfDNA fragment scores across healthy controls (green) and liver cancer patients (red) in the test set. Each line represents the distribution over all scores for a single patient's cfDNA sample, where the score for a single fragment is the probability that fragment originated from cancer. This data is patient groups 1 and 2.

I aggregate fragment scores to develop a patient-level risk score, which we then use to classify patients. In the held-out test set of unseen cfDNA samples taken from the same cohort as the training dataset, after integrating all fragment scores for an individual, all patients in the test cohort were correctly classified. In the second publicly available dataset of cfDNA samples from liver cancer patients and healthy controls, all patients were also correctly classified by the model, demonstrating its generalisability to new patient cohorts (Figure 4.5).

For the TAPS sequenced cfDNA samples, the model achieved an overall AUC of 0.75. Cell-free DNA samples from individuals with late-stage disease are correctly classified, but cfDNA samples from early stage disease are misclassified (Figure 4.5).

By obtaining classifications for each individual fragment in a patient's cfDNA population, we can begin to explore the relationship between methylation state, genomic sequence and fragmentation length of individual cfDNA fragments, and how this affects the predicted origin of the fragment. This was intended for the

TAPS-sequenced cfDNA samples from our pilot study, but model performance is too poor on these samples for the model fragment probabilities to be reliably interpreted as informative. To illustrate the potential utility of this approach in future improved iterations, I instead focus on model probabilities assigned to cancer cfDNA fragments in the first evaluation dataset which was correctly classified (where reads have average length 80bp). We can see that shorter fragments are assigned a lower cancer score on average. We can also see that as cfDNA fragment methylation rate decreases, the assigned probability that fragment originates from cancer increases.

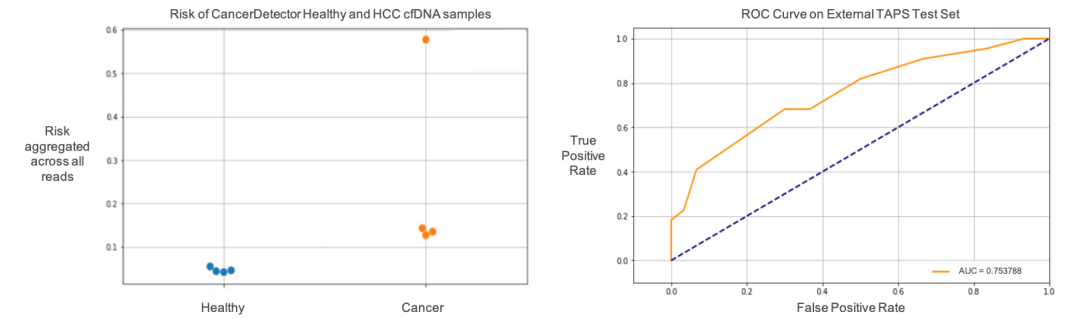


Figure 4.5: Left: patient risk scores for HCC and healthy cfDNA CancerDetector, taken by integrating all assigned model probabilities from individual fragments. Right: ROC curve generated for distinguishing TAPS cfDNA samples from HCC and non-cancer controls from patient-level risk scores, AUC = 0.75

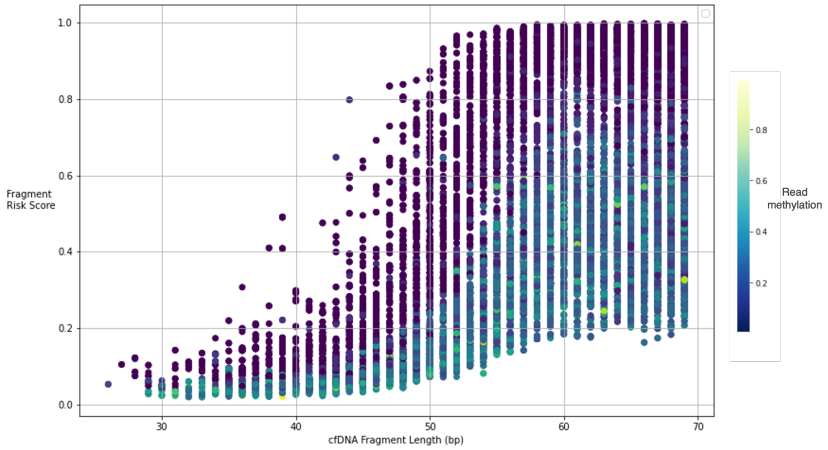


Figure 4.6: The distribution of cfDNA fragment risk scores stratified by fragment length, for a single liver cancer patient (downsampled for legibility). Methylation state (normalised by fragment length) is represented by marker color, where a lighter color means higher methylation.

4.6 Discussion

This study presents some initial attempts at maximising the use of multimodal information captured in new cfDNA sequencing techniques such as TAPS. Patient data is a scarce resource, and future diagnostic tests should be designed to make full use of available data, rather than discarding certain layers of information due to a lack of computational techniques for handling multiple concurrent modalities. I first show that combining information from multiple modalities can improve classification for more challenging tasks such as the three-way classifier discussed. The challenge here is not only distinguishing cancer samples from controls, but also distinguishing the two cancer types, HCC and PDAC from each other. Here we are effectively pooling knowledge from an ensemble of models trained on different modalities, and show these distinct models have complementary distinguishing powers in different areas.

A consistent limitation with this study and other cfDNA studies is we lack sufficient sample size to assess generalisability and thus ultimately utility of any

new method devised. This is a fundamental limitation of many recent studies, that can only be solved by collecting more data. However, in the fragment classifier approach, we partially mitigate this problem by not using precious cancer cfDNA samples for model training, but instead using tumour sequencing data which can be more readily obtained outside of trials than cfDNA. In this approach we train on an adjacent task of distinguish tumour data from healthy cfDNA, and the trained model is then applied to a novel objective by transfer learning.

A distinct but related problem is the difficulty of building supervised machine learning models that can learn interesting and useful representations of these biological datasets, given such small sample sizes. Learnt representations of biological molecules can be informative for downstream tasks, but rely on abundant training data to appropriately weight the input feature space. The fragment classifier a possible solution to this problem by swapping from treating each patient as a single data point, to treating each sequencing read as an individual data point. In making this switch, we go from having fewer than one hundred data points to having several billion data points for model training.

The availability of public tumour and cfDNA data was a limiting factor in this project. The fragment classifier developed here was trained on low-quality WGBS data, and accurately classified cfDNA molecules from the same cohort and other bisulfite sequencing experiments, but performed relatively poorly on TAPS-sequenced cfDNA. Certain experimental or sequencing differences between TAPS and WGBS may cause a data distribution shift that limit the generalisability of the trained model to this novel data type. For example, WGBS tumour samples were from single-end sequencing with average read-length of 80bp, whereas TAPS cfDNA samples were from paired-end sequencing with average read length of 150bp. The harsh reaction conditions associated with bisulfite sequencing may also have degraded sample quality to a greater extent than TAPS cfDNA samples.

It is also quite plausible that the poor performance on TAPS-sequenced cfDNA was due to the differing cancer stage distribution in this dataset. The correctly classified HCC samples from the TAPS samples were all from late stage disease, whilst early stage disease was consistently misclassified. By contrast the correctly classified WGBS samples all had late stage liver cancer. It is possible that cfDNA

from these early stage samples simply had insufficient circulating tumour DNA to be detectable and correctly classified in this model.

In addition to classifying cancer, we are seeking to better characterise the biology of cfDNA molecules. One potentially useful way of doing this is to examine the relationship learned in such a model between a fragment's sequence, methylation status or length, and the probability of originating from tumour or a healthy background. Intriguingly, the model developed here assigns a lower probability of originating from cancer to shorter, unmethylated fragments, which is somewhat in contradiction to the literature consensus. There are two plausible explanations for this observation. Firstly, although most studies report cancer-derived cfDNA fragments to be shorter on average, we and others have reported that ctDNA fragments can also be longer than expected, and in general what is most consistently observed is simply a change in fragment length distribution in cancer [59, 144]. Similarly, although hypomethylation is often reported in tumourigenesis, hypermethylation of CGIs is also a widely observed phenomenon in cancer development [69]. Secondly and perhaps more pragmatically, these represent fragment risk scores from a model that the results have demonstrated is not yet well calibrated, so it is quite possible that this model needs further tuning on additional training data, before we can reliably examine model fragment risk scores. Although instructive to highlight possible utility of such a model, I was not able to fully explore this relationship here, because of the poor performance of the trained model on TAPS-sequenced cfDNA samples of interest. This will remain an active question for future iterations of this model once more suitable training data is collected.

A clear limiting factor in this study was the availability of high quality training data, in the form of SAM aligned reads files for tumour samples. The pancreatic cancer cfDNA samples were not evaluated as we lacked pancreas tumour to train a similar model for this cancer type. Tumour data used for model training should be extracted using the same experimental and sequencing methodologies as the cfDNA we evaluate on, to minimise distribution shifts between the two. Distinguishing tumour cfDNA from the overwhelming background of fragments originating from healthy blood cells is a challenging task, and we can only expect to detect a proportionally small number of tumour fragments, as demonstrated in Figure 4.4. Tumour sequencing data should therefore also be high-coverage and minimally

degraded, to maximise the chances of detecting this rare molecule. The need for high quality tumour data that could aid models such as these is a principle motivation for the construction of a comprehensive tissue and tumour reference atlas, which is detailed in the next chapter.

Chapter 5

Constructing a reference atlas of tissue methylome and hydroxymethylome signatures

5.1 Abstract

Methylation and hydroxymethylation are key regulators of cell fate, but our understanding of the spatial distribution and interaction of these epigenetic marks is limited by a dearth of high quality data. In this chapter I describe the collation of a tissue reference atlas, containing genome-wide, CpG-resolution methylation and hydroxymethylation signatures for a range of healthy tissues and tumour types. In addition to providing new biological insights, this reference is a valuable resource for analyses relying on epigenetic information, such as cfDNA deconvolution and cfDNA fragment classifiers, where as demonstrated in previous chapters, genome-wide training data is vital yet scarce. Rather than relying on pre-defined regions for spatial stratification of this atlas, I develop a new segmentation algorithm to identify functional units of methylation and hydroxymethylation from the observed data. I adapt our previous DMR-calling method to these new spatial coordinates, and identify highly tissue-specific regions for both methylation and hydroxymethylation. These tissue-informative markers are enriched in tissue-specific gene expression and distinct histone modifications. Finally, I identify tumour-specific markers for both methylation and hydroxymethylation, thus uncovering a potentially distinct role for hmC in cancer. This reference atlas corroborates emerging evidence

of hydroxymethylation as a distinct epigenetic marker, and its potential utility as a novel cancer biomarker in training diagnostic models.

5.2 Introduction

The role of cis-acting epigenetics in determining healthy and malignant cell phenotypes is poorly characterised. This is due in part to a dearth of reference datasets that are uniformly processed and are therefore directly comparable. A single resource processed in a consistent manner allows us to identify trends across tissues and groups whilst minimising the risk of batch effects as a confounding variable. Methylation is known to be a key regulator of gene expression, and the genomic landscape of methylation is highly diverse. Some areas such as heterochromatin are densely methylated constitutively across tissues, other areas such as CpG Islands are constitutively demethylated, still other loci show variable, tissue-specific methylation, such as promoters of tissue-specific genes.

Some spatial details of this regulation are missing as most available data is from methylation arrays. This has led to a form of confirmation bias, where the covered sites in the array are more frequently studied so become increasingly linked to measured phenotypes, allowing their relevance to be reinforced. Meanwhile regions outside of these arrays are not studied and their functional significance remains unclear. At the time of starting this project, there were no publicly available genome-wide, CpG resolution methylation reference datasets available, aside from the previous reference atlas I constructed. This previous reference is formed of public low-coverage methylation data, so cannot be used to obtain high confidence read counts at individual CpG-resolution. I set out to rectify this by building a tissue reference atlas for methylation and hydroxymethylation across the entire genome at CpG-level resolution and high sequencing coverage. This atlas also include mC and hmC tumour data matching ten of the solid tissue types, and pre-cancerous inflammatory tissues for liver and pancreas.

Two atlases have recently been published since beginning this work, [91, 155], which reflects the interest in this resource for downstream methylation-based tasks such as tissue deconvolution. These atlases are a useful public resource, but they are still missing two key aspects: corresponding tumour samples, and CpG-level

genome-wide hydroxymethylation data. Including tumour samples in a reference is both informative for tumour epigenetics and useful for downstream tasks such as cfDNA deconvolution and fragment classification. Tumourigenesis is often associated with a dramatic reshaping of the epigenetic landscape. To establish the spatial details of these changes we need a matched healthy baseline to compare to. Including pre-cancerous inflammatory tissues for liver and pancreas provides an intermediate step in this progression. The presence of tumour samples in a reference atlas is invaluable for downstream tasks such as cell-free DNA deconvolution, as it allows us to distinguish tumour specific epigenetic signatures from tissue-specific patterns, as was seen for the liver-tumour specific contribution in HCC in Chapter 3.

The general consensus around methylation tissue-specificity, distribution around functional genomic elements and its alterations in cancer are actually observations of a mixture of mC and hmC combined, as they are derived from assays such as bisulfite sequencing that cannot distinguish mC and hmC. Less is known about the role and global distribution of hydroxymethylation: whether it is always anticorrelated with mC, or has a distinct profile concomitant with being a distinct epigenetic regulator [41, 42, 43]. Open questions remain about the functional differences between mC and hmC, and what information we lose when observing the mixture. For example, global hypomethylation is commonly observed in cancer: does this occur in both mC and hmC for all tissues? Does hydroxymethylation exhibit tissue-specific and cancer-specific effects? What is the spatial interplay of these two epigenetic marks in regulating gene expression, and do they operate over the same spatial scales? By sequencing mC and hmC separately for all tissues in this atlas, we can pick apart these two epigenetic marks, and identify trends specific to one chemical modification but not the other.

Having assembled this novel tissue reference, we can begin to investigate how mC and hmC are spatially distributed across the genome in relation to each other. Importantly, we can also ask the question: what is the appropriate functional unit at which to consider variation in methylation and hydroxymethylation? I previously relied on functional regions such as enhancers, promoters and transcription factor binding sites for reference-based cfDNA deconvolution. Whilst these genomic elements have been repeatedly shown to be useful partitions of the genome

for a variety of analyses, they are defined on public chromatin marker distributions, not on mC and hmC data. These elements are often fairly large, and do not necessarily represent the scale at which methylation and hydroxymethylation fluctuate. Further, reliance on predefined regions risks introducing bias into subsequent analysis, by assuming these are the appropriate functional units for mC and hmC. Given high coverage read counts across the genome at CpG resolution, we can instead take a data-driven approach to feature definition in mC and hmC space.

The processive, coordinated nature of cytosine methylation by DNA methyltransferases (DNMTs) means that adjacent CpG sites often have similar methylation states, leading to 'blocks' of local spatial correlation along the genome [159]. Recent evidence suggests these blocks, rather than individual CpGs, may be the relevant functional unit of methylation, creating a local methylation environment that shapes DNA accessibility and recruits DNA binding proteins to facilitate gene expression [160, 91]. We investigate this by segmenting base resolution methylomes and hydroxymethylomes of healthy and tumour tissues, uncovering the functional building blocks of these two epigenetic marks.

Segmenting the genome into spatially correlated blocks can be seen as a data compression exercise, effectively reducing the dimensionality of the input feature space from 29 million CpGs to much fewer functional building blocks of methylation. These blocks are treated as the new functional unit of analysis in downstream tasks such as identifying tissue-specific markers and tissue deconvolution of cell-free DNA, thus reducing the subsequent statistical burden on feature selection.

Genomic segmentation into spatially correlated blocks is a fairly new concept, so few algorithms exist. This is partly due to a previous lack of data: this reference atlas and a cell-type atlas published in 2023 [91] are among the first datasets with sufficient sequencing coverage at individual CpG level to warrant such exploration. More generally, sequence segmentation and change point analysis has a rich literature of methods, which we can borrow and adapt for genomic sequences. The objective of segmenting the genome is ultimately to find a globally optimal solution that still preserves biologically meaningful structure locally.

In this chapter I segment genome-wide mC and hmC data for all 116 tissue samples, covering 22 healthy tissue types and 10 solid cancer types. I first evaluate

an existing approach to methylome segmentation, and find it is not well suited to our hmC data across these scenarios. I then evaluate four new segmentation algorithms across a wide range of such scenarios. I find a hidden Markov model is the most robust segmentation method overall, and then use this algorithm to segment the entire methylome and hydroxymethylome.

The purpose of this new dataset is to be used as a public reference resource, containing tissue and tumour informative methylation and hydroxymethylation signatures. In the last part of this chapter, I therefore adapt the previously used DMR calling algorithm to identify tissue-specific differentially methylated blocks (DMBs) for both mC and hmC. I extend this to identify tumour-specific DMBs, and explore the possibility of pan-cancer markers in this dataset. I demonstrate the resulting DMBs capture tissue- and tumour-specific information, and validate them with orthogonal biological datasets.

5.3 Constructing a methylation and hydroxymethylation reference atlas

Note: this was a collaborative project, where M. Inoue performed the experiments necessary to collect the data, and I led the analysis with collaboration from J. Cheng

To build a methylation and hydroxymethylation reference, we first needed to source tissue samples for sequencing. Our aim was to build as comprehensive a reference atlas as possible. We prioritised breadth over depth, such that most of the major organ systems were represented, at the expense of not including individual cell types for each tissue. This was also informed by my previous work on cfDNA tissue deconvolution and the broader literature: I needed to ensure the major organ systems and contributors to cfDNA such as liver and immune cell types were included in this reference. We also elected to include tumours for each tissue type where possible. This was an important missing piece from our previous reference, and indeed all other published reference methylation datasets. We ultimately collected 122 tissue samples, which were sequenced in our lab using TAPSbeta and CAPS protocols. This generated genome-wide methylation and hydroxymethylation data for each of these 122 tissue samples. The schematic

below summarises our dataset. Full details for tissue atlas data collection, patient metadata and sample processing can be found in the Appendix.

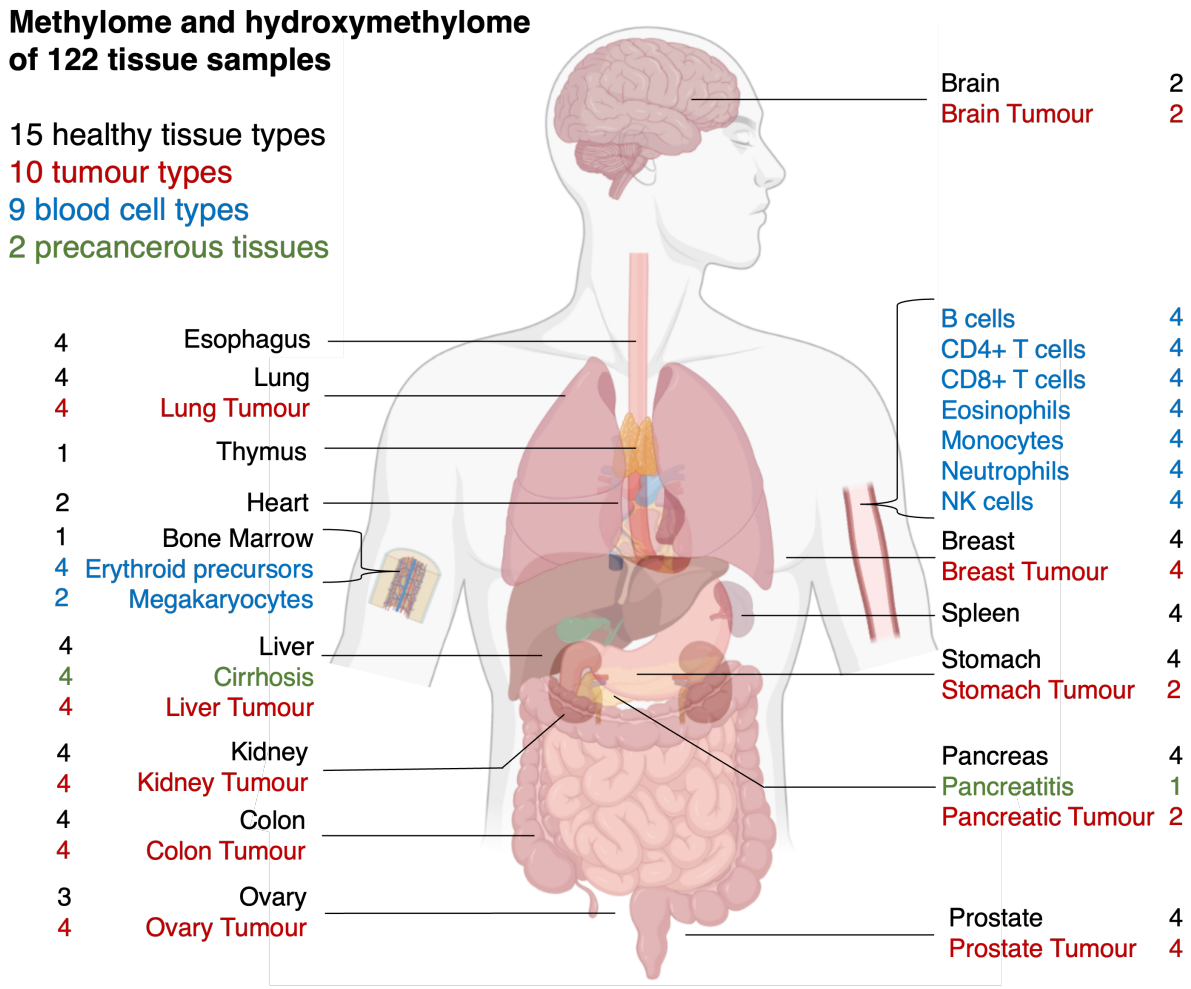


Figure 5.1: Schematic of all tissue samples collected for our reference atlas

5.3.1 Tissue sample quality control and clustering

Before incorporating samples into our reference atlas, a number of quality control (QC) checks were necessary to judge suitability of each tissue sample for inclusion. The following criteria was used for each sample:

1. Sufficient depth of coverage

2. Even read coverage across the genome

3. Expected methylation ratio distribution

4. No batch effect, clustering near other samples from the same tissue

All tissue samples had a high depth of coverage, with a global mean of 30.76x across all TAPSbeta samples and 30.12x across all CAPS samples, as illustrated in Figure 5.2. The mean methylation ratios for hmC and mC across all samples and all 28 million CpGs is shown in Figure 5.3 [1]

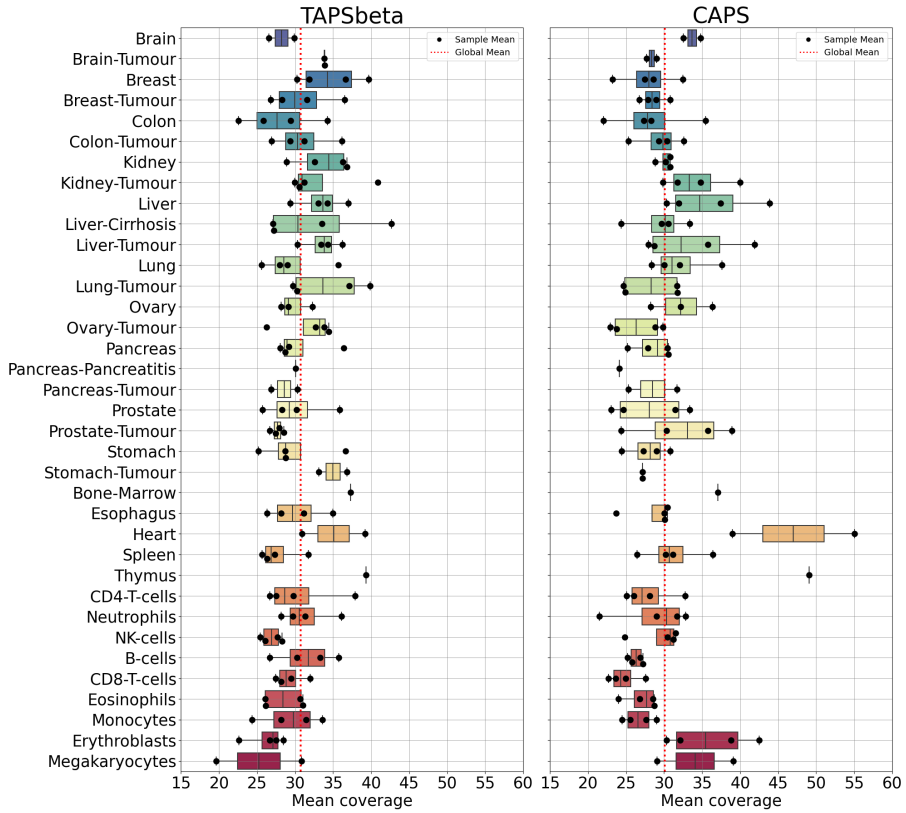


Figure 5.2: Mean coverage across all CpGs for all samples in the two experiments, TAPSbeta and CAPS.

[1]Note: the terms erythroblasts and erythroid precursors are used interchangeably in this Chapter; they refer to the same blood cell type in this reference atlas

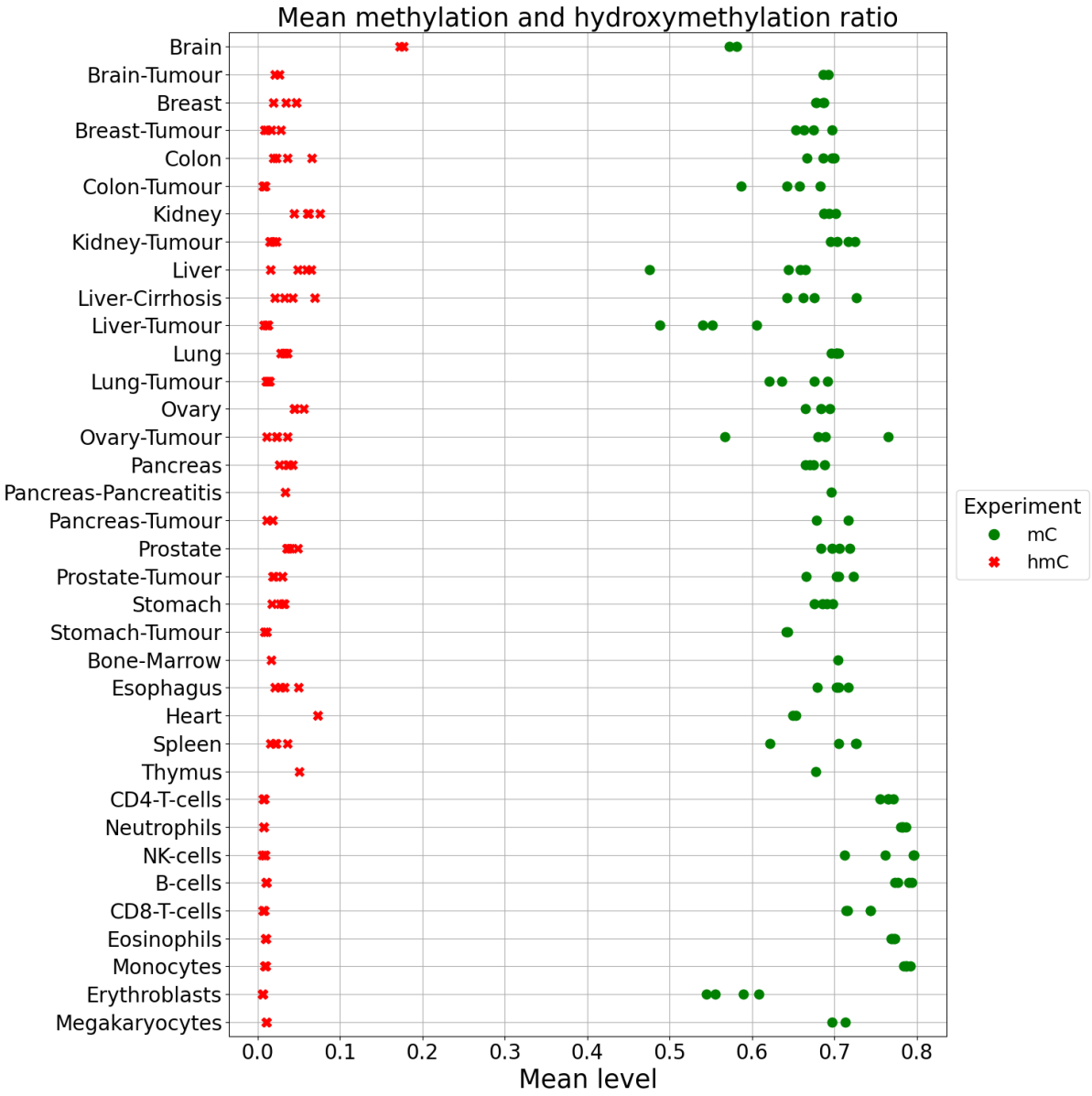


Figure 5.3: Mean methylation and hydroxymethylation ratios for all samples

Six samples were removed on the basis of uneven coverage or a highly unusual methylation ratio. These included the only tissue replicates of thymus and bone marrow. After removing these six sample outliers, the final dataset was comprised of 116 tissue samples, and 22 healthy tissues, 10 tumour types and 2 tissues in a pre-cancerous inflamed state. The QC steps taken to filter samples in this atlas

are described in Appendix 1. These remaining tissue samples demonstrated clear within group clustering. A t-SNE embedding [150] of the final tissue reference atlas with methylation ratio averaged across Enhancers is shown in Figure 5.4.

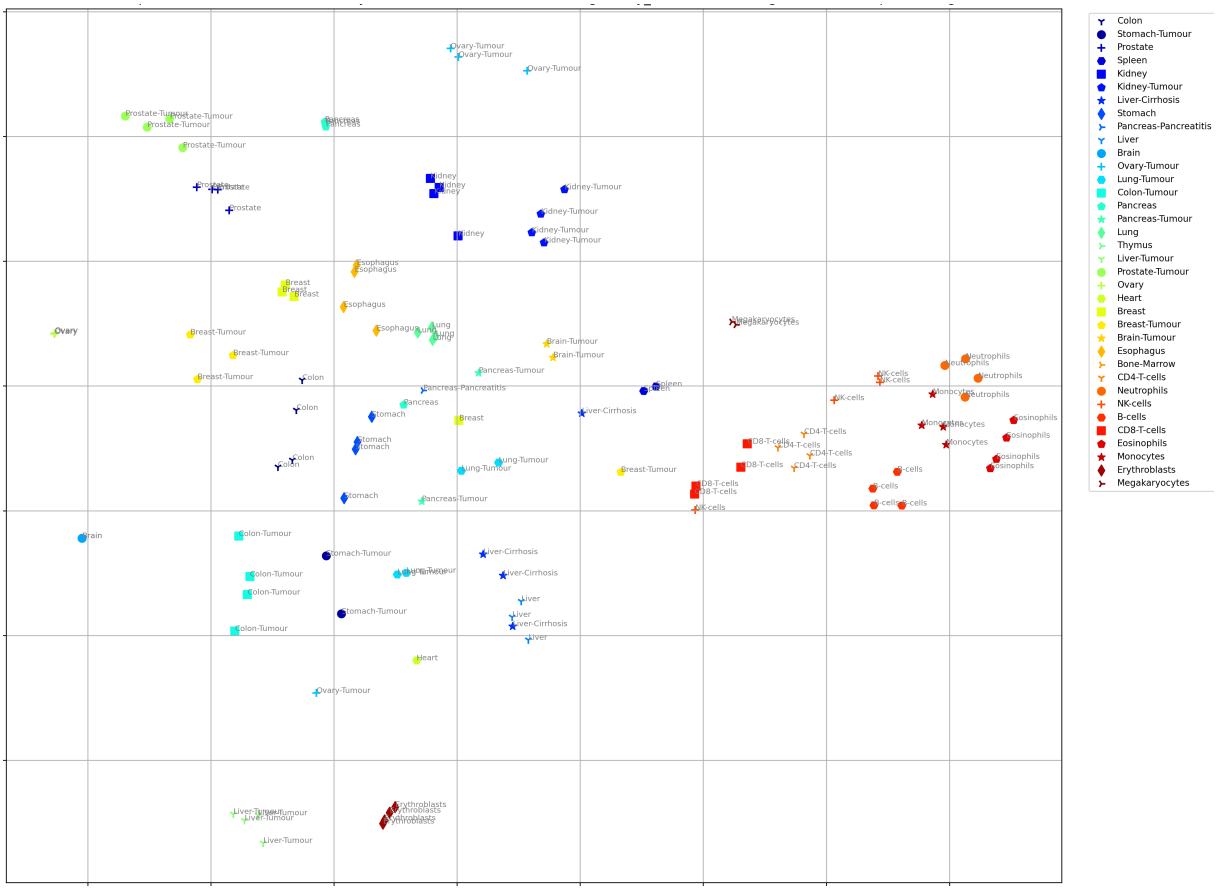


Figure 5.4: t-SNE projection of final reference atlas after removing outliers

I compared mean methylation and hydroxymethylation levels across all CpGs for healthy tissue and tumour samples. Some unexpected differences emerge between methylation and hydroxymethylation levels of healthy samples compared to corresponding tumours. For example, we generally see global hypomethylation in tumour samples, as reported in literature [1], but this is not always the case. Kidney tumour methylation levels are mean increased compared to healthy, as are brain, ovary and pancreas to some extent, as shown in Figure 5.5. By contrast we see a consistent decrease in mean global hydroxymethylation levels for all tumour

types. Previous epigenetic sequencing methods such as WGBS cannot distinguish mC and hmC, so we would observe just a single modification level which is actually a combination of both methylation and hydroxymethylation. The distinct trends seen in these two separate epigenetic marks would be masked in such experiments. To illustrate this, I combined mC and hmC marks to generate an overall modification level similar to what would be observed in these experiments, using MLML [161] (Figure 5.5, bottom panel). Such details are already a good indication of the useful additional information gleaned from separating mC and hmC in our reference atlas.

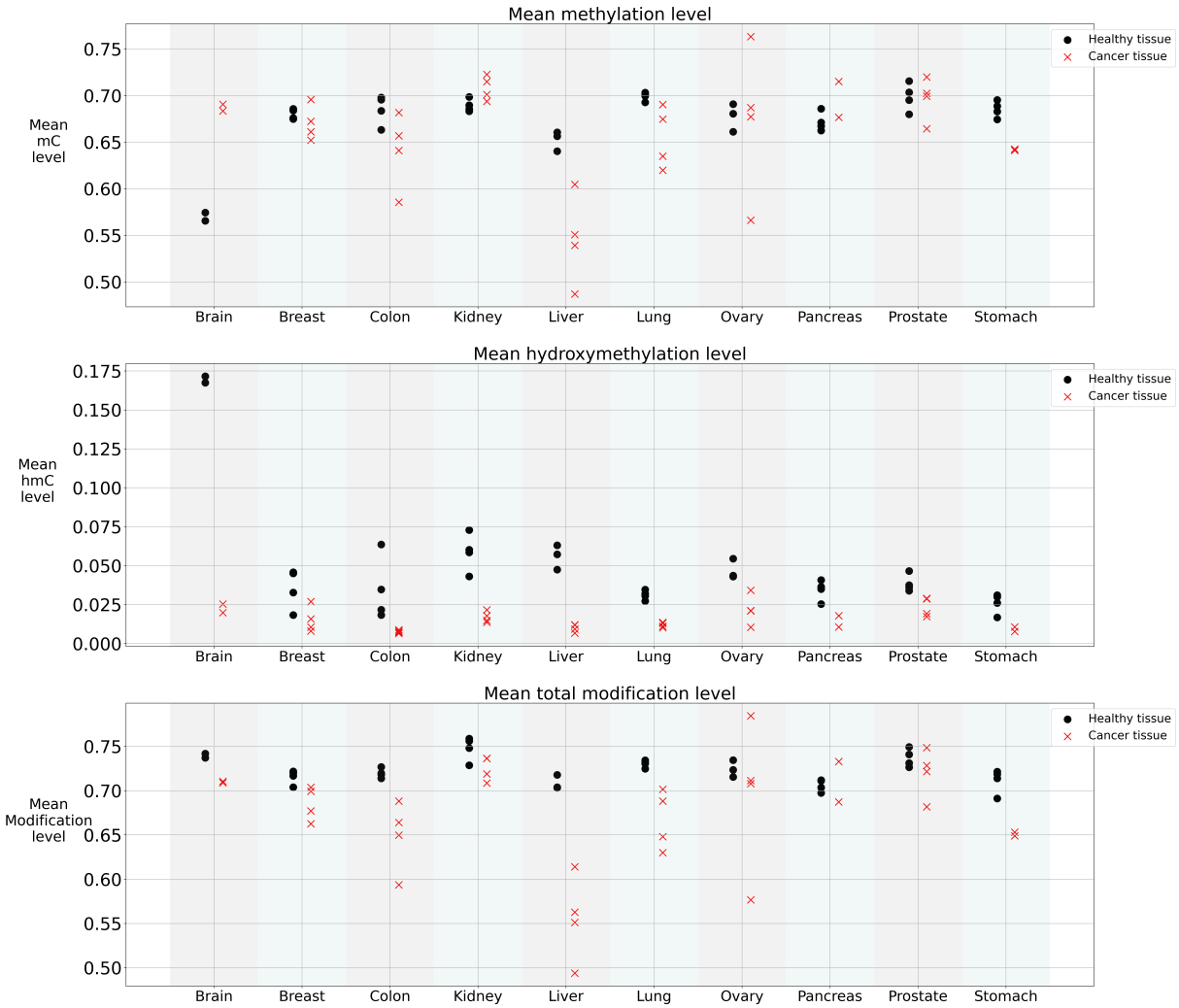


Figure 5.5: Mean global levels of mC (top) and hmC (middle) across all CpGs in our reference. Bottom panel shows overall modification as calculated using MLML.

5.4 Segmenting methylation and hydroxymethylation into functional building blocks

Having finalised the set of high quality tissue samples in this reference atlas, I sought to spatially segment methylation and hydroxymethylation profiles of all tissues simultaneously, in order to ascertain the functional building blocks of these epigenetic marks.

5.4.1 Evaluating existing segmentation methods

A methylome segmentation method based on change point analysis was published earlier this year [91], and before devising new methods I sought to apply this method to mC and hmC data. I repeated their segmentation and filtering schemes, and obtained a similar result as reported in their paper for methylation, with approximately 1.5M mC blocks after filtering. I found clear tissue-specific differentially methylated blocks in our data using their method. However, the size distribution of hmC blocks was very different, on average 10x larger, and I found few tissue-specific hmC blocks, nor could I cluster tissue samples into sensible groups. I therefore concluded that this method was unsuitable as a data compression algorithm for hmC, and sought to devise my own method.

5.4.2 Benchmarking methods on synthetic data

I devised and optimised segmentation methods by benchmarking on synthetic data. I simulated methylation data for 30 tissues by sampling from a beta-binomial distribution around true tissue derived means along a stretch of CpGs, as illustrated in Figure 5.6. As we process along CpGs in this example, there are clear areas of locally similar methylation state across all tissues. These areas of spatially correlated methylation are followed by distinct changes in the overall methylation state. The problem statement is simple: how can we segment the genome so as to optimally group together regions of correlated methylation, and draw boundaries where the local methylation state changes.

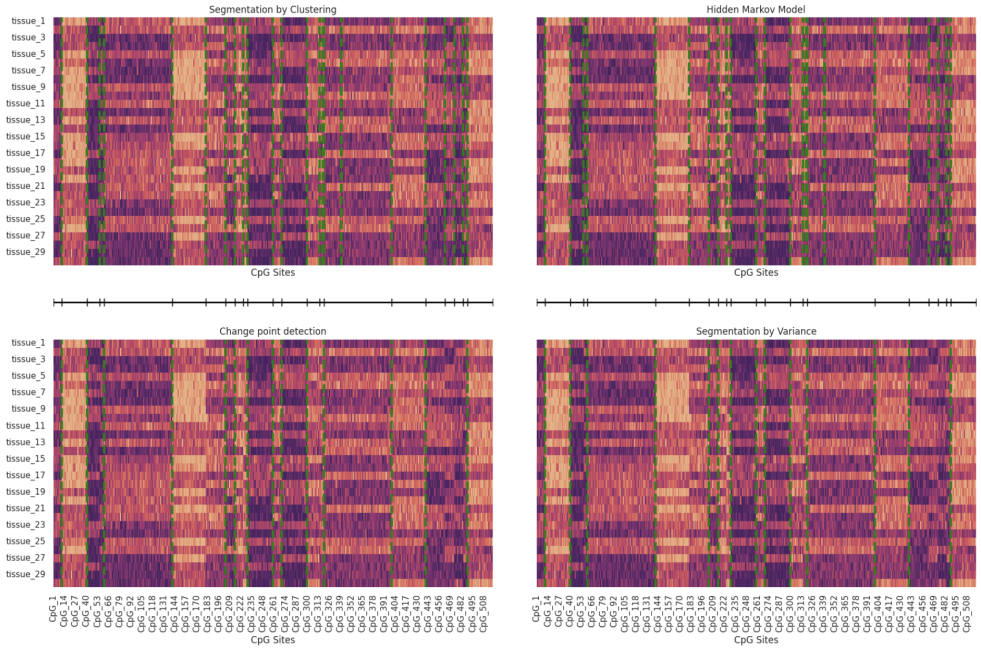


Figure 5.6: An example of the simulated methylated data used for benchmarking algorithms. The black lines indicates the true positions of the blocks in the simulated data, and the dashed green lines indicate the assigned boundaries found by each of the four segmentation algorithms.

Method	Summary
Change Point	Change point detection algorithms attempt a global minimisation of a cost function, based on the log-likelihood of the data. The Pruned Exact Linear Time (PELT, Killick 2012) algorithm implemented here penalises additional change points with a penalty, and finds the segmentation amongst all possible segmentations that minimises the cost function. Implemented using the ruptures library in Python.
Variance-Based	Segments the methylome by processing along each CpG, calculating cumulative variance at each step. The cumulative variance should only decrease as we process along a homogeneous block with the same methylation state. A significant increase in variance above a threshold value is then taken as a segmentation boundary. The position is recorded, and the cumulative variance starts over in this new block.
Clustering	Applies hierarchical clustering to the methylation data to segment the genome into blocks based on similarity[2]. The resulting clustering is discarded, but the boundaries between clusters are taken as block boundaries.
hidden Markov model	Implements a Gaussian HMM to segment the genome based on transitions between underlying hidden states inferred from the methylation data.

Table 5.1: Summary of segmentation methods implemented for methylation data.

The segmentation methods I evaluated are summarised in Table 5.1. Figure 5.6 demonstrates the slightly differing results of the four segmentation algorithms applied to a single simulated region. For each method, I simulated a range of data scenarios and captured the trade-off between precision and recall of each method using the F1-score. This allowed us to optimise for a segmentation method that is conservative in boundary assignment yet still correctly identifies the majority of true simulated boundaries. Benchmarking results are summarised in Figure 5.7. I found a hidden Markov model with two hidden states had best overall performance across a range of scenarios. The clustering approach and the "PELT" method of change point detection both performed better in some cases, but were less consistent overall, and less robust to simulated noise. I also tested the time

[2]Inspiration for this method came from our collaborators M.D and B.S.B whom we gratefully acknowledge.

complexity of each method, a key consideration when applying this algorithm across the entire width of the genome, and found that the HMM scaled best to increasing size of input data.

5.4.3 Testing segmentation algorithms on tissue methylation data

I next evaluated these segmentation methods on real methylation data. I used base resolution genome-wide methylome data from 22 healthy tissues (77 individual tissue samples), generated by TAPSbeta and CAPS sequencing. I again evaluated the four segmentation algorithms on real data, fine-tuning the parameters of each method. There are no ground truth "correct" block positions when segmenting real data, so I evaluated overall segmentation of a genomic locus using two unsupervised metrics: a modified Bayesian Information Criteria (BIC), and Silhouette Score. The modified BIC score replaces likelihood with within-group variance of methylation, a measurement of the homogeneity of methylation within individual blocks, as follows:

$$\text{BIC} = 2\sum_{i=1}^{k} \bar{\sigma}_i^2 + k \cdot \ln(n)$$

where $\bar{\sigma}_i^2$ is the mean variance of the i-th segment, k is the number of inferred boundaries, and n is the number of CpGs. Silhouette score measures quality of within group clustering resulting from the assigned segmentation boundaries, and is defined in the Background section.

The genomic landscape of methylation is highly diverse: I tested each segmentation method across a wide range of such scenarios and found that a HMM was the most robust segmentation method overall. I repeated the same process for hmC and found the same result. Segmentations around some tissue-specific genes for mC and hmC are shown in Figure 5.8.

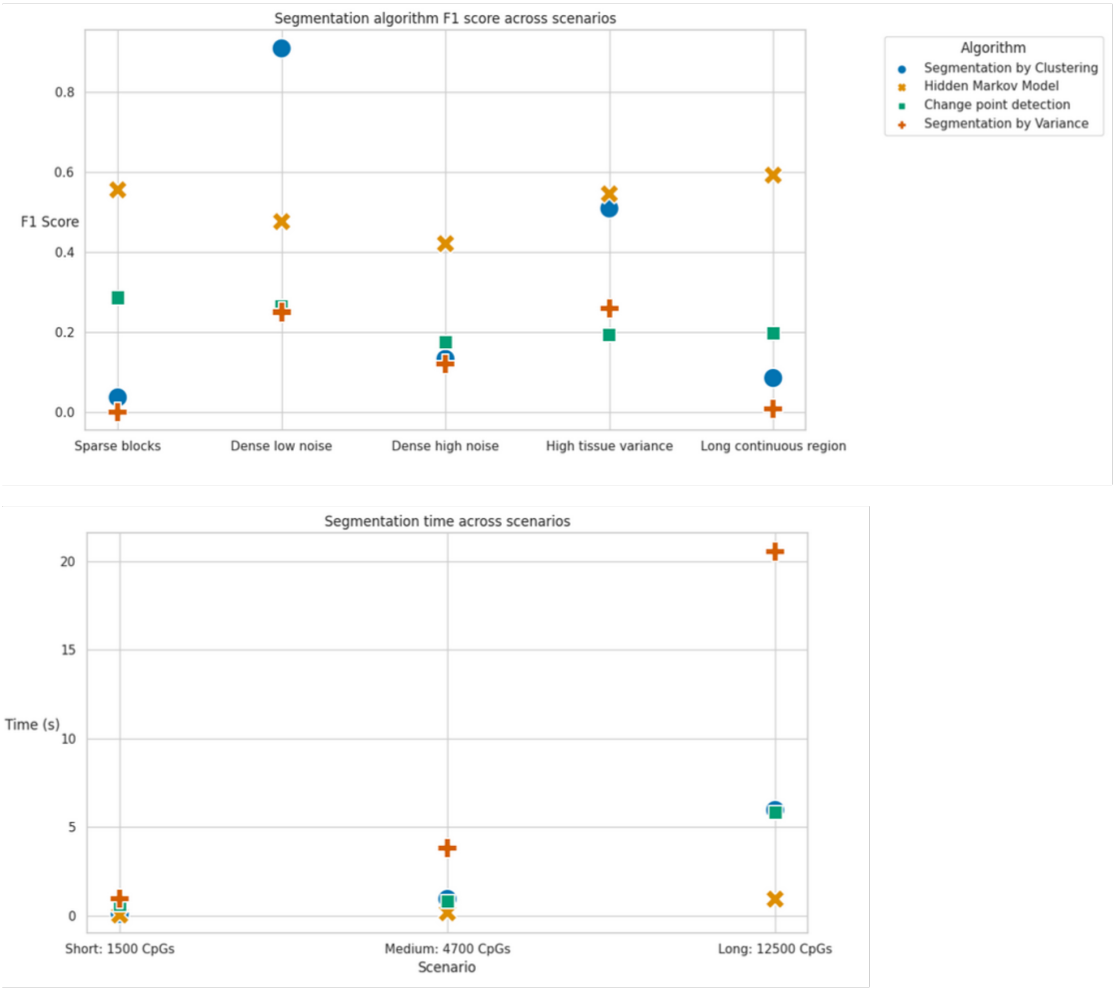


Figure 5.7: Comparing performance of the various segmentation algorithms. Left: F1 score of each segmentation algorithm, comparing the number of true simulated blocks correctly identified against the number of false positives across different simulated scenarios. Right: the time taken for each algorithm to run on different data scales.

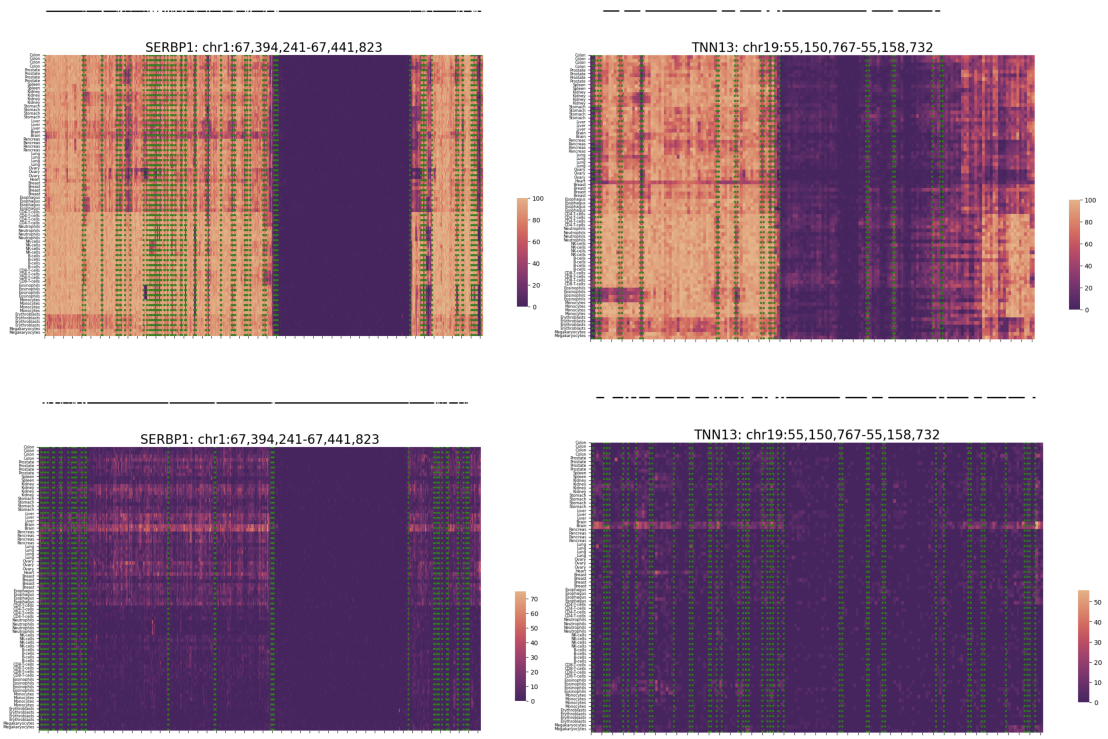


Figure 5.8: HMM segmentation of mC (top) and hmC (bottom) around the SERBP1 and TNN13 genes.

There is no obvious biological rational for picking any particular number of hidden states in the HMM. I therefore chose the optimal number of hidden states for the HMM segmentation method by evaluating a range of parameter values with a composite score, combining BIC and Silhouette score. Generally speaking, increasing the number of hidden states in the HMM model led to a higher frequency of smaller blocks being called at any given genomic locus. Increasing this parameter is akin to increasing the sensitivity of the segmentation, allowing smaller fluctuations to be detected at the cost of being less robust to noise. This is demonstrated for gene A1BG in Figure 5.9. I repeated this analysis across different epigenetic landscapes and across numerous tissue-specific genes, and generally found a HMM with two hidden states was optimal, as demonstrated in 5.9. Repeating this process for hmC data, I also found two hidden states was optimal for hydroxymethylation segmentation.

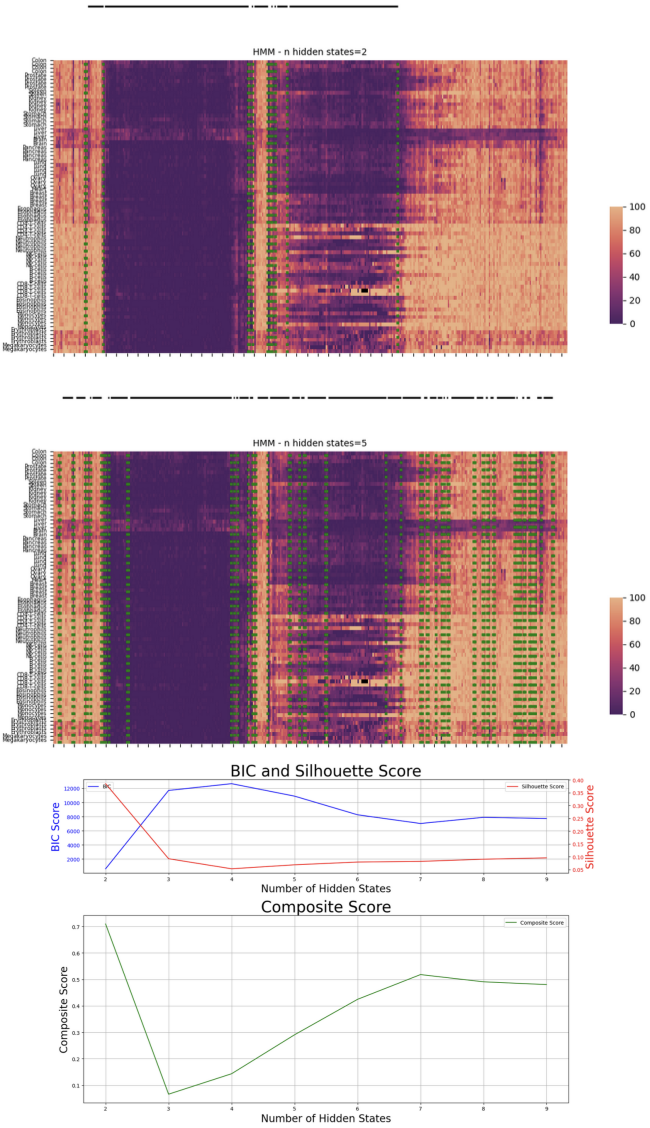


Figure 5.9: Comparison of different HMM parameter values over a single region surrounding the A1BG gene. The upper two panels show the resulting segmentation for two distinct values for HMM hidden states. The lower panel shows the BIC and silhouette scores for these parameter values, and the composite score from combining the two with equal weighting.

5.4.4 Segmenting mC and hmC across all tissue samples

After optimising the segmentation algorithm, I proceeded to segment the entire methylome and hydroxymethylome of all samples, using a hidden Markov model with 2 hidden states. The resulting segmentation identified 4,046,528 blocks for mC and 5,173,896 for hmC, and their distributions of length and number of CpGs are showing in Figure 5.10. A large number of small blocks are found due to local noise; for downstream analysis I filtered by blocks containing 4 or more CpGs. Hydroxymethylation levels in brain are very distinct and vary approximately 10 times more than all other tissues. I was therefore concerned that including brain tissue may overly bias the resulting hmC segmentation, so repeated the segmentation after removing it. The overall block distribution did not change significantly, so I elected to include it in order to retain the additional information this distinctive tissue provides.

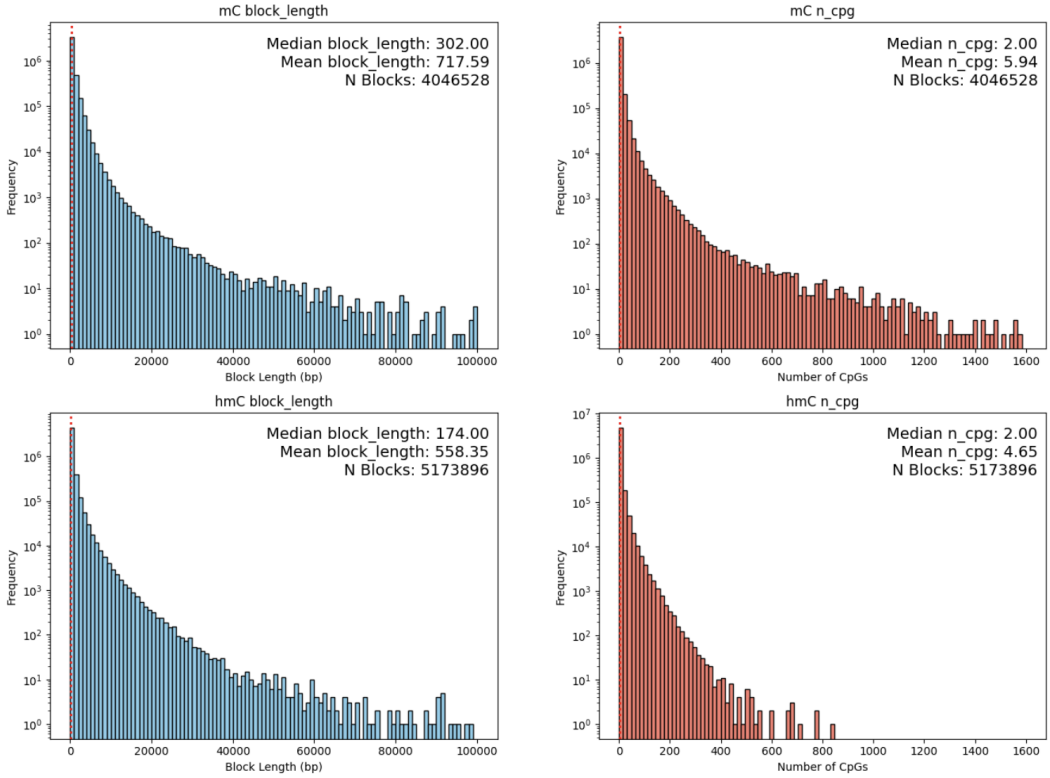


Figure 5.10: Median block length (left), and median number of CpGs per block (right) for mC and hmC segmentation (top and bottom respectively)

5.4.5 Segmenting the cancer methylome and hydroxymethylome

I separately segmented mC and hmC data for the 34 tumour samples from 10 solid cancer types. Overall I found a larger number of smaller blocks when segmenting tumour samples compared to healthy tissue samples: the mean block length changed from 718bp down to 613bp for mC, and from 558bp down to 469bp in hmC. A typical change for mC is illustrated below for the liver tumour specific gene TOP2A in Figure 5.11.

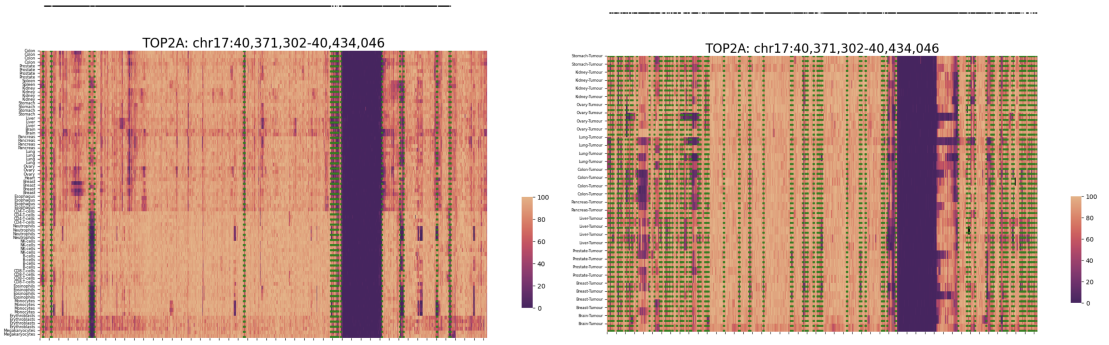


Figure 5.11: A comparison of segmenting the same region using the same algorithm, but on tumour methylation data (right) compared to healthy samples (left).

5.4.6 Unsupervised clustering of tissue samples over blocks

To test whether the resulting blocks were informative, I performed hierarchical clustering on all tissue samples over the top 5% of highest variance blocks, replicating the approach used in [91]. Both mC and hmC blocks clustered well into tissue groups, and had high clustering ARI and NMI scores (defined in Chapter 3), as shown in Figure 5.12. These tissue clustering scores were considerably higher than those observed for the pre-existing segmentation method implemented on our data.

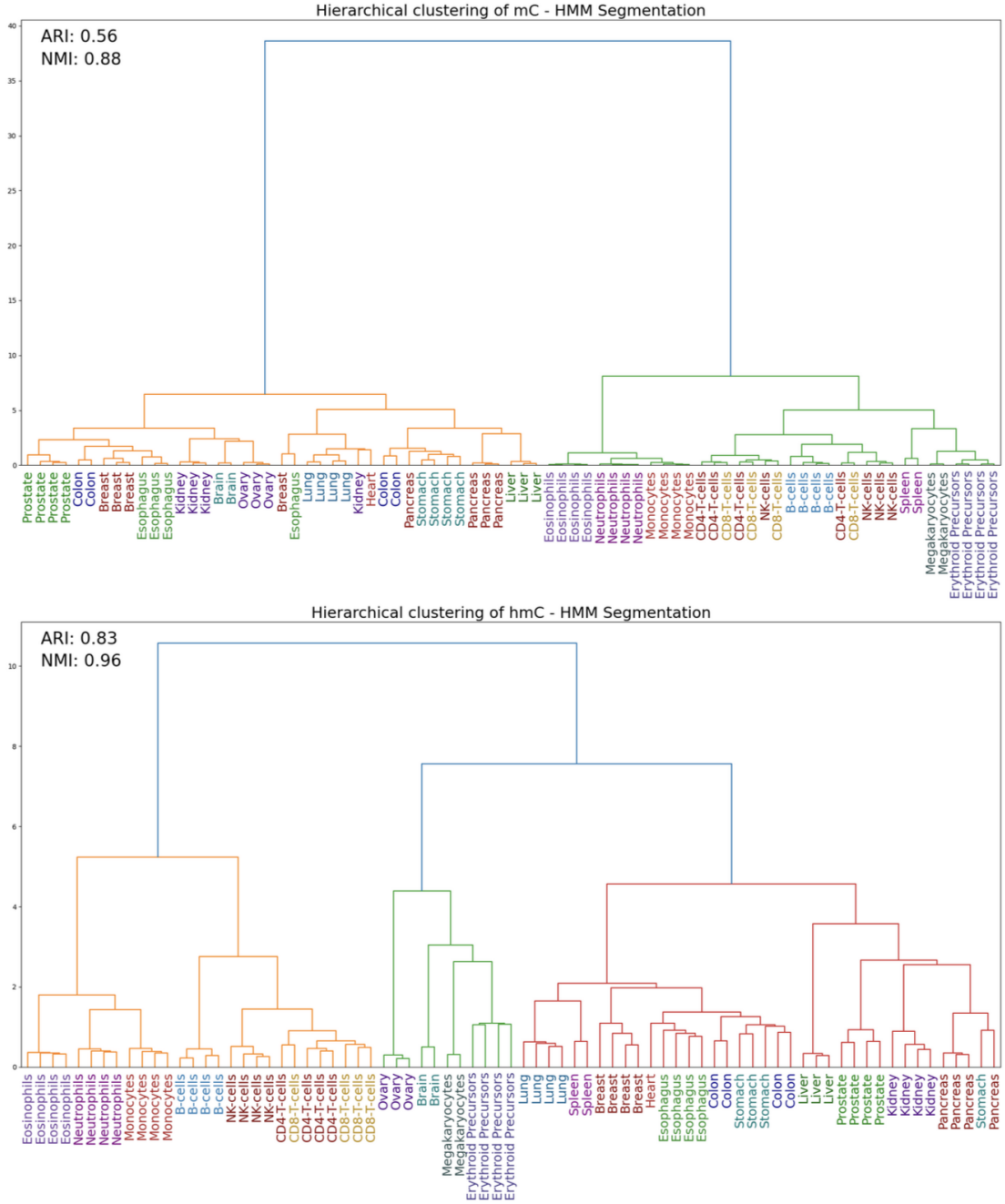


Figure 5.12: Dendrogram clustering of tissue samples using methylation and hydroxymethylation blocks. NMI and ARI scores indicate the quality of the clustering.

5.5 Identifying tissue-specific differentially methylated and hydroxymethylated blocks

Having segmented the genome into blocks, we are interested in obtaining differentially methylated and hydroxymethylated blocks that are highly tissue-specific: that can be used to distinguish one tissue in our reference from all others. I tested several of the DMR methods shown in Table 2.1. These published methods were largely not applicable for our use case, for three main reasons. Firstly, most of these methods were defined for pairwise case control settings, as defined above. Methods such as RADMeth[116], Metilene [118], DMRFinder [119] and DMRcate[120] implement custom region aggregation algorithms, performing local smoothing and merging adjacent differentially methylated CpGs into regions. These custom pairwise segmentations are not applicable for our use case, as each pairwise comparison for each tissue is defined on distinct non-overlapping genomic features. I have already defined a global segmentation across the genome for all samples, and use these underlying features throughout. Having defined this new genomic feature set, these methods generally implement fairly straightforward statistical tests for differential methylation, such as a paired t-test or Wald test. I adapted the relevant statistical tests from these methods, skipping the first step.

Secondly, the case control setting these methods were built for often assumes numerous biological replicates exist for each group, so that more robust statistical methods that involve parameterising a distribution can be applied. This is not always the case in our data. The small sample size of the target group negates the applicability of most statistical tests: mean tissue group size is four, but can be as low as one for some tissue groups such as Heart. For example a t-test assumes normality between groups, which is a contentious assumption with a sample size of < 30; for a sample size of two this assumption cannot be made. Furthermore, considering the small sample size of the target group and the variable effect size, we often would not have sufficient statistical power to correctly reject the null hypothesis that the observed sample means are sampled from the same underlying population distribution.

Finally, the code to implement these algorithms is often inaccessible, as it is packaged into a software suite which requires commitment to their data formatting

pipeline to access the DMR finding method. These methods are built for methylation array data or WGBS, and often the data formatting assumptions they require are not suitable for TAPSbeta or CAPS data.

5.5.1 Identifying tissue-specific differentially methylated blocks

After evaluating a range of published methods, I found that the most suitable approach to finding tissue-specific DMBs is a delta threshold method, similar to the method used for the public WGBS reference atlas in Chapter 3. One important improvement I made was instead of filtering regions by median difference between target and background group, I used custom percentiles for the target and background groups; these groups have largely different sample sizes, so should have a tailored tolerance for outliers. For hypomethylated DMBs, I use the 75th percentile of the target group and the 2.5th percentile of the background group. For hypermethylated DMBs, I use the 25th percentile of the target group and the 97.5th percentile of the background group. This is a small change, simply moving from the 50th percentile of both groups to new specific percentiles for each group, as illustrated in Figure 5.13. This "soft thresholding" approach finds regions with large methylation deltas between the target groups, but still allows for some outliers in each group. I set the percentile thresholds such that if there is one outlier in the target group (average size of 4), it does not overly influence the calculated value. There are many more samples (typically more than 70) in the background group, so use the 97.5th percentile, which still allows for a few outliers but enforces fairly uniform levels in the background group. These percentiles worked well in another recent tissue methylation atlas publication [91], which is why I adapted them here. For brevity, the difference in methylation (or hydroxymethylation) values between these two percentiles will often subsequently be referred to as "delta quantiles" or "delta quants".

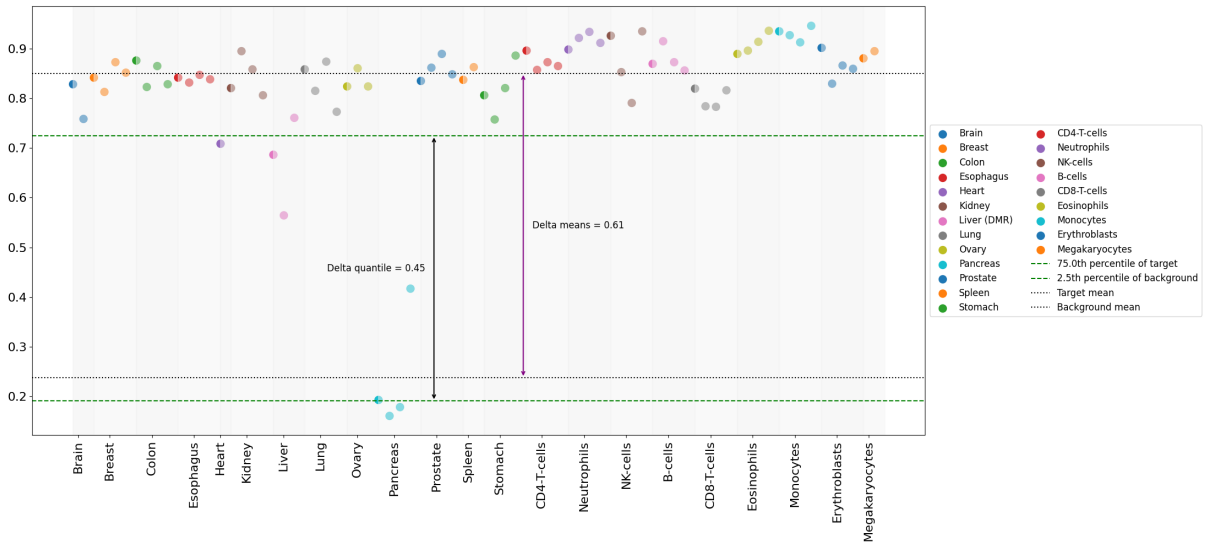


Figure 5.13: Illustration of new DMB finding approach for a real Pancreas DMB. The new appproach ranks regions by the methylation delta between the 75th percentile of the target group and the 2.5th percentile of the background group, for hypomethylated DMBs. My previous approach used the methylation delta between group medians.

We then rank regions by their delta values between these two percentiles, and choose a threshold cut-off number of DMBs depending on the specific context. Figure 5.14 shows the top 300 methylation DMBs by this criteria, for each of the healthy tissue types in our atlas. Figures 5.14 and 5.15 show the top 300 tissue-specific DMBs by this criteria, for mC and hmC respectively, for each of the healthy tissue types in our atlas. Clear tissue-specific DMBs were seen for mC and hmC, and less than 2% of these DMBs were overlapping between mC and hmC. Not surprisingly given their background levels, hypomethylated DMBs were most clear for methylation, whilst hyper DMBs were most distinct for hydroxymethylation.

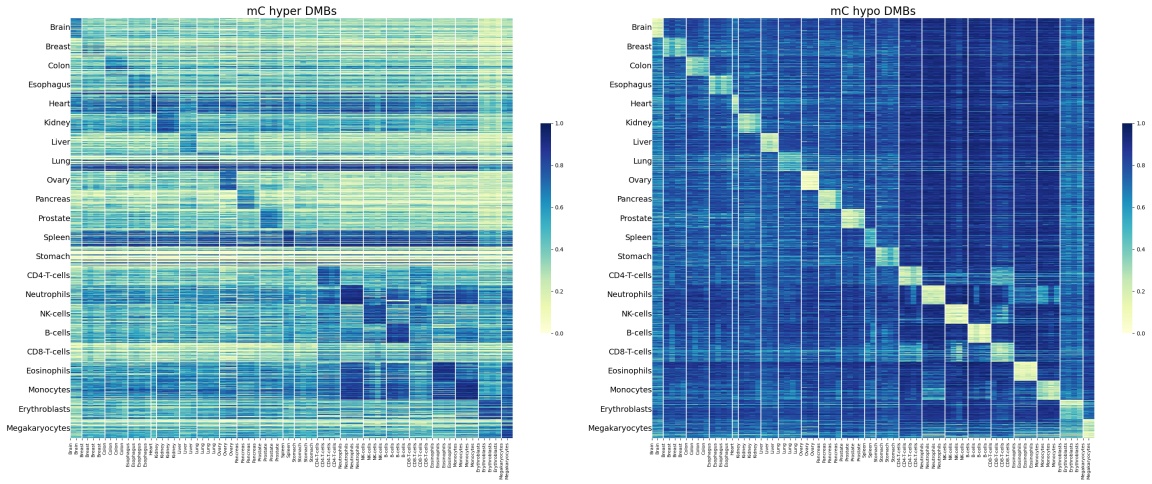


Figure 5.14: Top 300 hypermethylated (left) and hypomethylated (right) mC DMBs for each healthy tissue in our tissue atlas

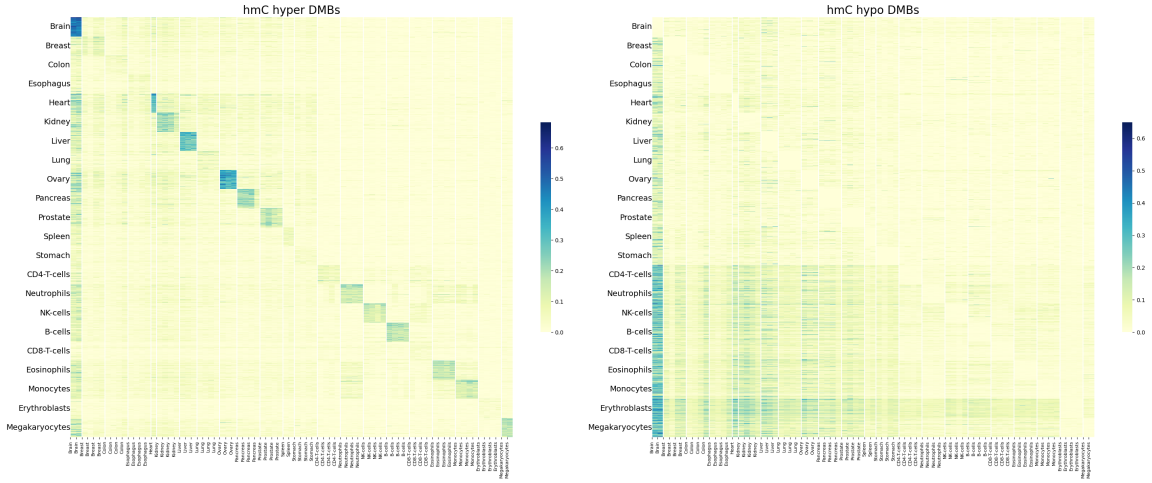


Figure 5.15: Top 300 hypermethylated (left) and hypomethylated (right) hmC DMBs for each healthy tissue in our tissue atlas

Across the literature, there are no consistent rules for selecting a cut-off point beyond which a candidate region is no longer be considered a DMR, thus distinguishing DMRs from non-DMRs; this is of course linked to the lack of an exact definition of DMR. Most methods rely on some kind of statistical test, then filter regions by a p-value cutoff corrected for multiple testing. However, these methods generally rank all regions by their test statistic of choice, then keep the number

of regions required as an adjustable parameter in the model, leaving the user to choose a sensible cutoff for their particular context and application.

I explored three options for pinpointing a cutoff to obtain an optimal set of DMBs.

1. Use a manually set hard threshold, such as methylation delta between means > 0.1

2. Devise a statistical test, and filter DMRs by FDR adjusted p-value significance.

3. Devise some metrics to measure 'optimal' number of DMRs to include for our purpose

I investigated implementing various cutoffs using statistical tests, such as t-test, Welch's t-test (for unequal variance and unequal sample size), and a p-value derived from permutation testing (repeated bootstraps to estimate the data distribution). Figure 5.16 shows all healthy tissue mC DMBs where the delta between the two quantiles is greater than zero, coloured by Welch's t-test significance. This shows that t-test significance and the delta quantiles filtering metric largely correspond and can be used interchangeably. Ultimately it seemed more pragmatic to simply take the top n DMBs satisfying a robust and easily interpretable statistic such as delta between quantiles, rather than retro-fitting a statistical test and all the concomitant assumptions, solely to obtain a significance score.

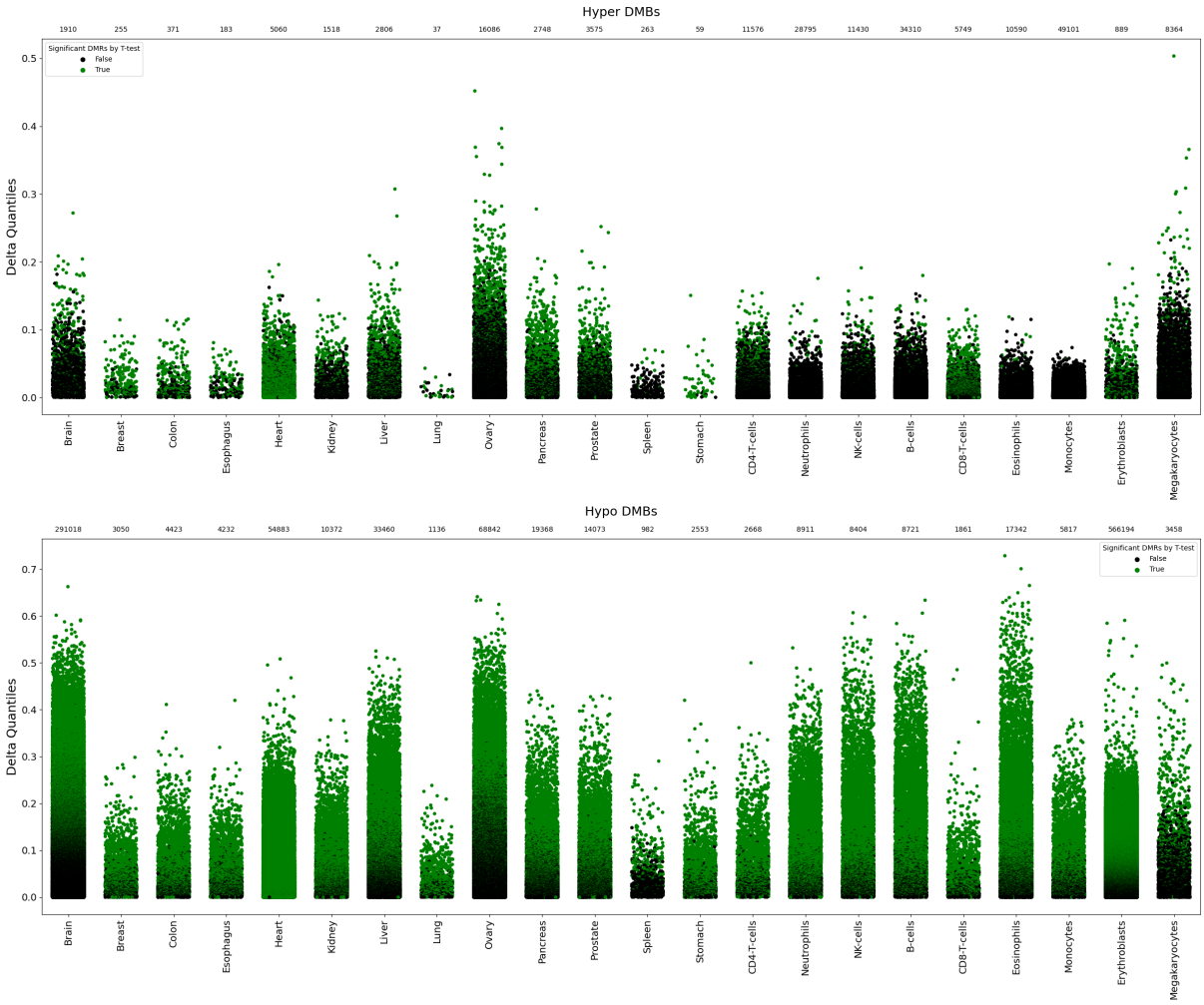


Figure 5.16: All TAPSbeta DMBs for healthy tissues, split into hyper DMBs (top) and hypo DMBs (bottom), and coloured by whether or not each DMB is assigned a Bonferroni corrected significant p-value.

5.5.2 Downstream interpretation of tissue-specific DMBs

To further characterise these tissue-specific DMBs, I first investigated how they are distributed across genes, relative to functional elements such as promoters, transcription start sites and the gene body. I took the subset of tissue-specific DMBs for each tissue and calculated their distribution across the entire gene functional region for tissue-specific genes. This distribution is calculated by first dividing all tissue-specific genes into 40 equal-length windows, plus 40 250-bp bins either side

of the gene body, then counting how many DMBs fall into each window, normalised by total number of genes. I aggregated results across all tissues, to make a meta-gene plot as shown in Figure 5.17. I again focused on the most prominent DMB type for each epigenetic mark, so used hyper DMBs for hmC and hypo DMBs for mC. Tissue-specific hypomethylated DMBs were dramatically enriched around the TSS region, with a steady drop-off in frequency throughout the gene body. By contrast, tissue-specific hyper hydroxymethylated DMBs were distributed fairly evenly across the gene body, and depleted around the TSS region.

I also calculated enrichment of mC hypo DMBs and hmC hyper DMBs in each functional region. These functional regions have a very diverse range of sizes, and enrichment score is affected by amount of genome captured in each region. I therefore calculated an odds ratio of enrichment for each functional region by permutation testing: 100 random permutations give the feature enrichment expected by chance, then compare the observed enrichment to this baseline to obtain an odds ratio for DMB feature enrichment. Hypo mC DMBs and hyper hmC DMBs were both primarily enriched enhancers, with odds ratios of 2.4 and 1.39 respectively.

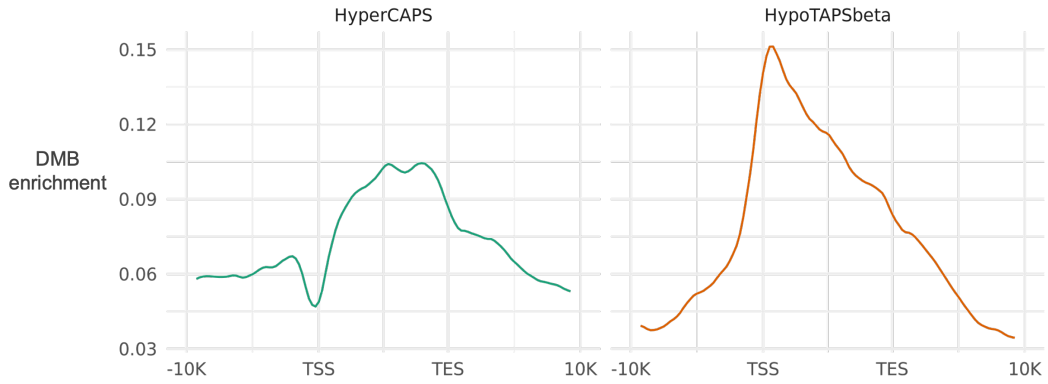


Figure 5.17: Metagene plot calculated across all genes and all tissues, showing the distribution of DMBs across the gene body. The y-axis shows the number of DMBs linked to tissue-specific genes, normalised by the number of tissue-specific genes. This is shown separately for hyper-hydroxymethylated DMBs (left) and hypomethylated DMBs (right). The x-axis indicates the transcription start site (TSS) and transcription end site (TES), plus 10-kilobases in either direction. The region between TSS and TES is considered the gene body.

I then sought to validate these tissue-specific DMBs, to determine whether they were true tissue informative markers, or simply an artefact of this dataset. I took the top 300 hypo DMBs for mC and hyper DMBs for hmC, and assigned each DMB to its nearest proximal gene on the genome by base-pair distance from DMB to gene TSS. Note, this assignment carries several assumptions, such as that each DMB has a cis-acting role in its immediate vicinity, but was the simplest available with the fewest assumptions, and has been used recently in the literature [91]. Having assigned each DMB to a proximal gene, I tested whether these genes showed tissue-specific expression. I calculated enrichment (using Fisher's exact test) in a list of genes showing tissue-specific expression, downloaded using EnrichR [162] and originally taken from GTEx and PanglaoDB [163, 164]. For example, for the 300 liver-specific DMBs linked to genes, I calculated the enrichment in terms within tissue-specific gene lists in GTEx. Overall I found candidate genes for each list of DMBs were highly enriched in gene lists relating to that tissue, as shown in Figure 5.18.

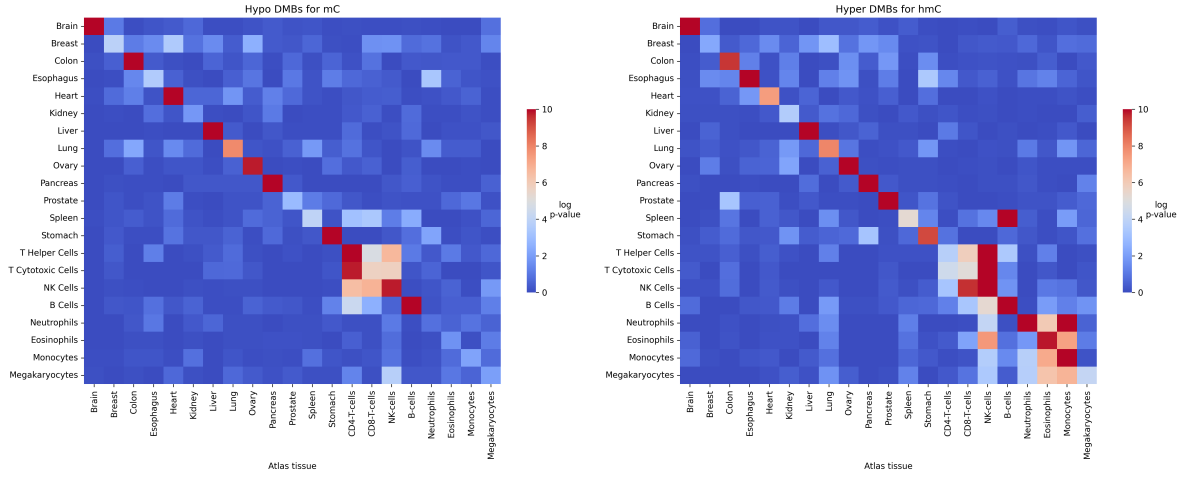


Figure 5.18: Enrichment of top 300 DMB-linked genes in tissue-specific genes downloaded from EnrichR, with EnrichR tissues on y-axis and atlas tissues on x-axis. The same method is shown for mC hypo DMBs (left) and hmC hyper DMBs (right). For tissue-specific genes from solid tissues, the original dataset was GTEx Tissues V8 2023; for immune cells the original dataset was PanglaoDB Augmented 2021.

Finally, I investigated enrichment of tissue-specific DMBs in histone modifi-

cations. Chromatin accessibility is known to vary as a function of histone modification density, and these modifications are taken as the *de facto* markers of different chromatin states [145]. For example, acetylation of H3K27 (H3K27Ac) is commonly linked to open chromatin and active expression, whilst methylation of H3K4 (H3K4Me1) is commonly associated with enhancer regions. The hypomethylated mC DMBs I identified were enriched in H3K27Ac and H3K4Me1 in a highly-tissue-specific manner, across all types in our atlas. The hmC hyper DMBs were also tissue-specifically enriched in H3K27Ac and H3K4Me1 for some tissues in our atlas, but not all. Examples of the general trend of this tissue-specific histone enrichment are shown in Figure 5.19, for two representative tissues.

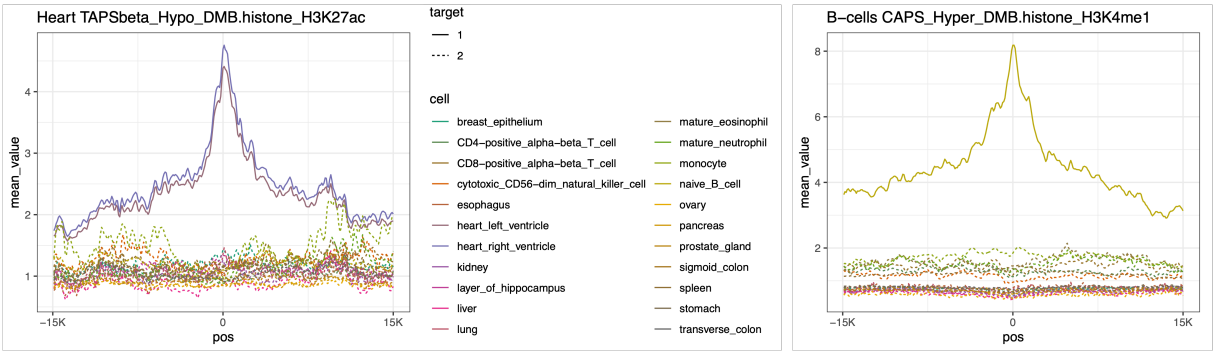


Figure 5.19: Enrichment between hypo mC DMBs and H3K27Ac (left), and hyper hmC DMBs (right) for a variety of tissues. The profile of histone markers across many other tissues are shown for comparison. Histone enrichment for the target tissue is shown with a solid line, and other non-target tissues are shown with dotted lines. For example, this indicates that around heart-specific DMBs from our tissue atlas, we observe enrichment in H3K27Ac histone marks for heart tissue, but not for other tissues.

5.5.3 Identifying cancer-specific differentially methylated blocks

It was not clear to what extent tumour-specific DMBs would be obtainable in this reference atlas, considering both the fairly small sample size and the inherent variabiity of tumour samples. However, using the same DMB calling algorithm, I identified clear tumour-specific hypo DMBs for mC (Figure 5.20). To an extent, hyper-hydroxymethylated regions are also discernible for hmC (Figure 5.21). In-

terestingly, those tumour types that demonstrated clearest hyper-DMBs for hmC are the same tumour types demonstrating a distinct pattern in global levels between mC and hmC, as shown in Figure 5.5. Cirrhosis and pancreatitis tissue samples are not shown here, but they exhibited intermediate methylation values for these liver and pancreas tumour DMBs respectively.

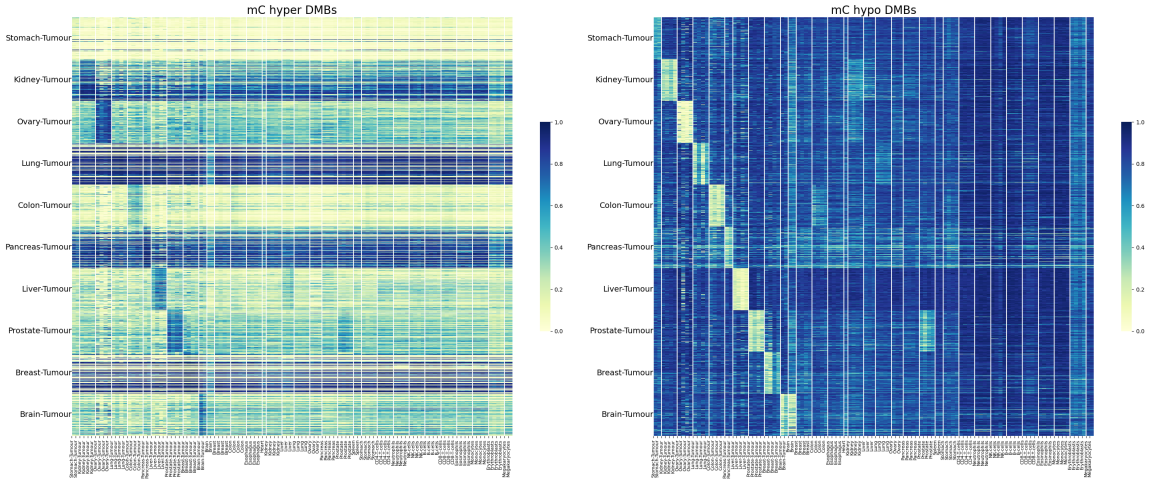


Figure 5.20: Hyper and hypomethylated tumour specific DMBs for methylation

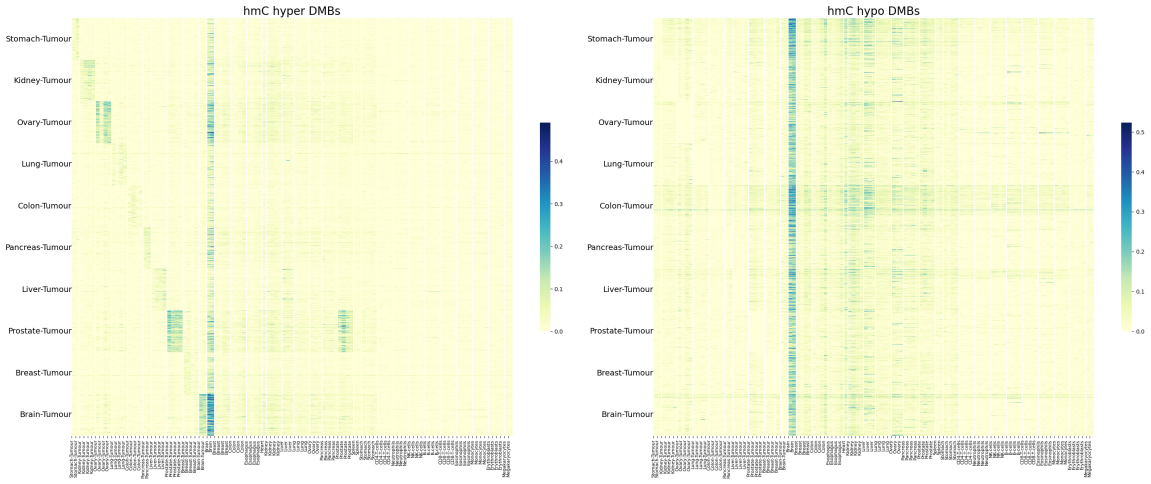


Figure 5.21: Hyper and hypomethylated tumour specific DMBs for hydroxymethylation

As for healthy DMBs, if we set a hard threshold on DMB filtering criteria, we

see different numbers of DMBs for different tumour types. We also see a tumour-specific proclivity for either hyper or hypo DMBs. Figure 5.22 shows all tumour methylation DMBs with a delta quantiles value greater than zero. DMBs are coloured by p-value significance, as assessed by Bonferroni corrected Welch's t-test. If we use this hard threshold, different tumour types have widely varying numbers of DMBs. For example liver tumour has the vast majority of hypo DMBs (more than 400,000), but proportionally much fewer hyper DMBs; conversely, prostate tumour has proportionally many more hyper DMBs that pass this hard threshold.

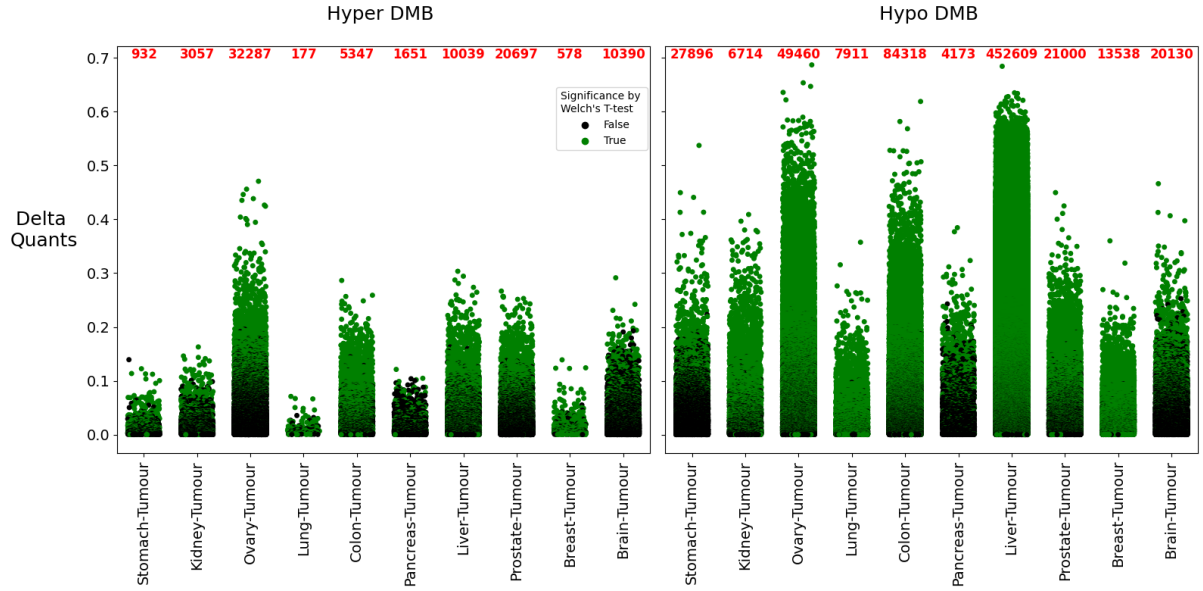


Figure 5.22: DMBs for each tumour type ranked by value of the methylation delta between assigned quantiles, coloured by p-value significance in Bonferroni corrected Welch's t-test. The red text shows the total number of DMBs.

As for healthy DMBs, I validated these mC and hmC tumour DMBs by assigning them to their nearest gene and examining enrichment of tumour-specific genes. We observe clear tumour-specific enrichment, as shown in Figure 5.23.

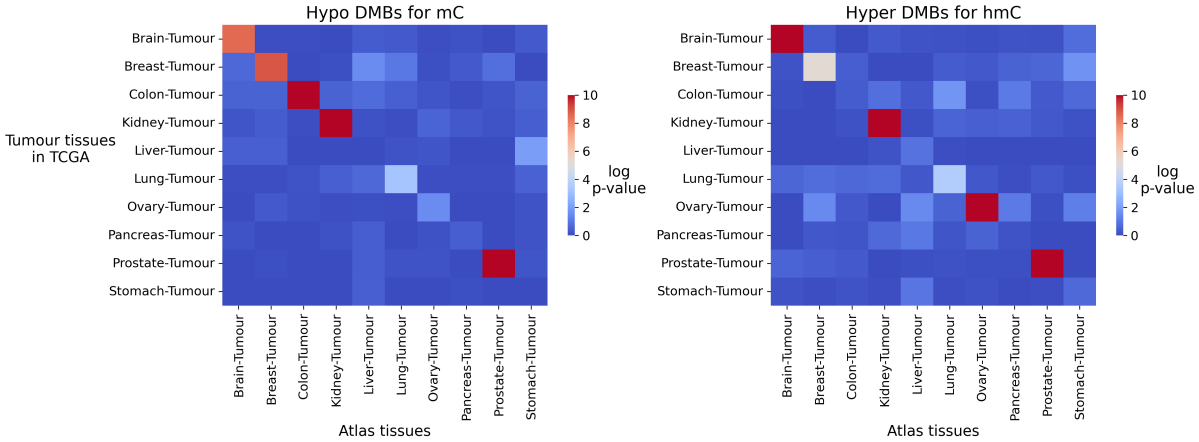


Figure 5.23: Enrichment of tumour DMB-related genes within up-regulated genes in tumour samples. Left shows enrichment for hypo mC DMBs and right shows the same for hyper hmC DMBs.

5.5.4 Deconvolution of cell-free DNA using tissue informative markers

Having collated this reference atlas and identified differentially methylated blocks for healthy and tumour tissues, I investigated using this new data for tissue deconvolution of the previous cfDNA samples. I took the top 300 mC hypo DMBs defined for all healthy and tumour types in this atlas, and as in my method in Chapter 3 added additional pairwise DMBs to distinguish correlated tissues. I intersected non-cancer and HCC cfDNA samples on these DMBs, and used the deep learning approach I previously developed for deconvolution, re-trained on simulated cfDNA mixtures by sampling from this new reference atlas. Importantly, the cfDNA samples were sequenced using previous TAPS sequencing [93] which did not distinguish between mC and hmC, but I use mC-specific DMBs as a reference for this task. This is not a perfect comparison, and in future iterations I will compare cfDNA methylation alone with this mC reference, and likewise for hydroxymethylation separately. Nonetheless, we observe a broadly similar tissue contribution to previous results, with blood cell types dominant and an increased liver tumour contribution in HCC.

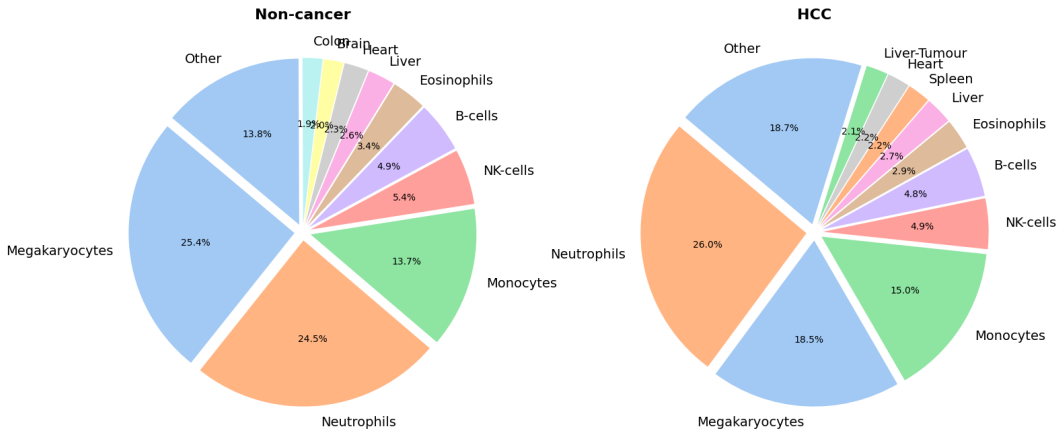


Figure 5.24: Predicted tissue contribution in HCC and non-cancer controls, using deep neural network deconvolution method re-trained on simulated cfDNA mixtures from this new reference atlas.

5.5.5 Searching for pan-cancer DMBs

Having identified tumour-specific hypomethylated regions, and to some extent tumour specific hyper-hydroxymethylated regions, I next asked the question: can we identify any DMBs in mC or hmC that are shared across all tumour types. This can be investigated in both a top-down and bottom-up approach. The top-down approach involves grouping all tumour samples together and all healthy samples together, and trying to identify DMBs between the two groups using any standard case versus control DMR finding method. The bottom-up approach involves identifying DMBs between each tumour and a background of all healthy samples in turn, then looking for common DMBs shared across each of these comparisons.

Figure 5.25 shows number of shared DMBs for each tumour type, after filtering DMBs by Welch's t-test significance. This number is normalised by the total number of DMBs for that tumour type. It was necessary to calculate this as a proportion so as to avoid bias introduced by the widely varying total number of DMBs for each tumour type: otherwise, over-represented tissues like liver would dominate shared DMBs across all tumour types, simply because it has an order of

magnitude more hypo DMBs than other tissues. It is interesting to observe that the tumours with the highest proportions of shared DMBs are often those where we see secondary metastates occurring, for example colon and liver, or breast and lung.

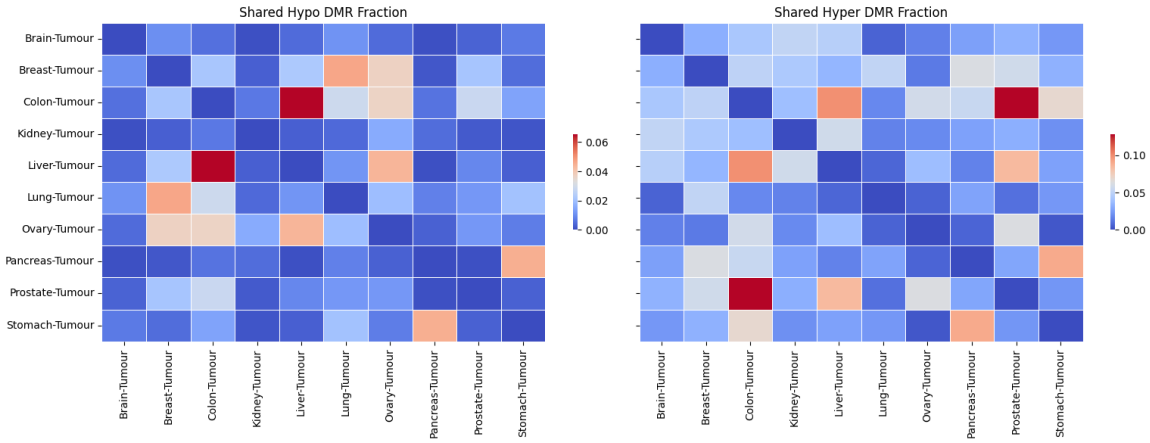


Figure 5.25: Proportion of shared DMBs for each tumour type

There were no pan-cancer DMBs with a clear difference between all tumour samples and all healthy samples, for either methylation or hydroxymethylation datasets. However, using the bottom-up method, I identified 49 DMBs that were shared between 7 or more tumour types. These regions can separate tumour samples from healthy fairly well in a simple PCA, as shown in Figure 5.26. Interestingly, all four DMBs shared between nine tumour types were all hyper DMBs, as were five out of seven of the DMBs shared between eight tumour types. Some examples of these DMBs are shown in Figure 5.27.

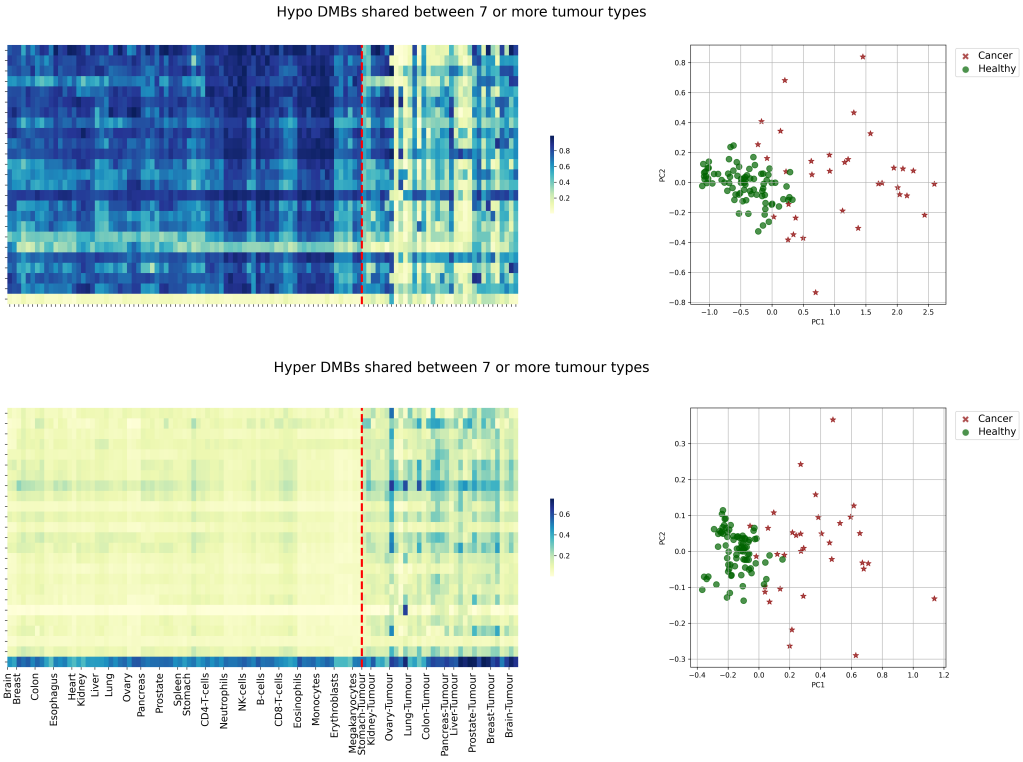


Figure 5.26: Proportion of shared DMBs for each tumour type

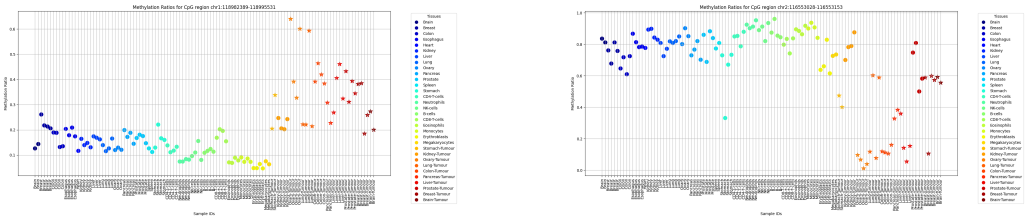


Figure 5.27: Methylation ratios of all samples in a hyper DMB and a hypo DMB shared between 8 tumour types. Tissue types classed as containing this DMR are starred.

5.5.6 Automated feature selection methods for tissue-specific DMRs

In many aspects of computational data analysis, methods have improved drastically by replacing manual feature selection steps with automated feature selection,

thus connecting more steps in the analysis pipeline with a differentiable loss function. This was one of the core innovations in deep learning, which first became prominent in computer vision: instead of handcrafting features based on an image's pixel statistics, we simply feed in the entire image into the model [165]. Each feature (pixel) in the data input is connected (via many non-linear activation functions) to a differentiable cost function, which learns to place a high weight on important features in the model input, and a low weight on irrelevant features. This approach has entered the medical domain, first with the replacement of 'radiomics' style manual feature selection methods for CT scans with large convolutional neural networks [166]. More recently, this approach has been successfully applied to biological sequences, with Transformer models performing automated feature extraction across amino-acid sequence space in AlphaFold [102]. Transformers have also been successfully applied to the genome sequence, where methods such as Enformer extract DNA motif features across a span of 100kb to effectively predict gene expression from DNA sequence alone [104].

Automated feature extraction has not yet been successfully applied to DMR finding, so I devised and evaluated some approaches here. The basic premise of automated feature selection is to replace the reliance on metrics and heuristics with a single differentiable cost function. This cost function should reduce given more informative features. A supervised learning model can thus take iterative steps towards identifying important features by learning coefficients that minimise this cost function.

I first attempted to identify DMRs using supervised learning methods, classifying one tissue versus rest. For example, to identify Liver-specific DMBs, I built a binary classifier to predict tissue labels, hoping the learnt model weights would point to Liver-specific DMRs. It was necessary to evaluate this classifier using leave-one-out cross-validation, with pooling of common DMRs shared across all iterations. However, the small sample size of each individual tissue group (average size of 4) constrains classifier performance here, and I was not able to accurately classify tissues. We are of course not interested in the classifier's prediction labels, except as a form of validation: if the classifier accuracy is low, then it is unlikely to be identifying tissue-specific regions. I also tried a multi-class classifier with 22 different classes equal to the number of healthy tissue groups, but accuracy of

this classifier assessed through k-fold cross-validation was also not good enough to continue.

I subsequently tried this classifier approach to identify pan-cancer DMBs. Here sample size is slightly higher, with 34 tumour samples and 77 healthy samples. Using a RandomForest binary classifier, I was able to accurately distinguish samples and thus identify cancer informative regions, which could separate tumour samples from healthy on PCA fairly well (Figure 5.28).

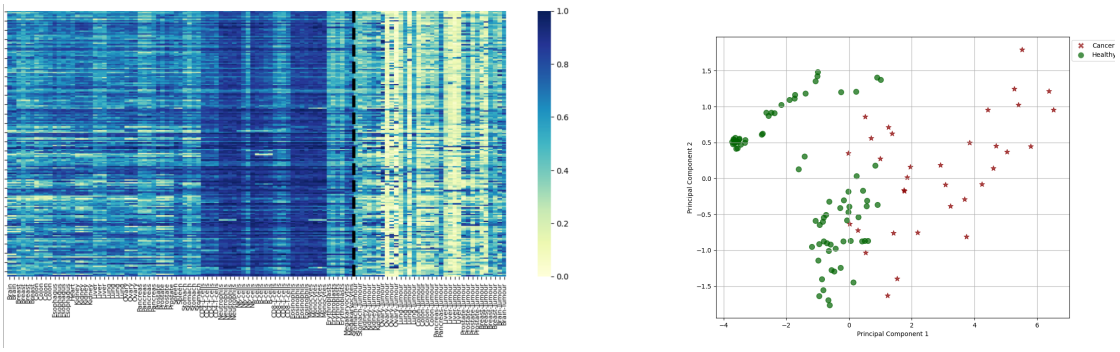


Figure 5.28: Left: Heatmap showing methylation values for pan-cancer features identified by RF classifier. Right: PCA demonstrating separation of samples on these genomic features.

5.5.7 Non-negative matrix factorisation to identify tissue-specific DMBs

Non-negative matrix factorisation uses a two-step optimisation method to identify the underlying signatures and their relative contributions to a given sample or collection of samples. It can therefore be used as a distinct, differentiable method for identifying tissue-specific DMBs: we can take the highest weighted components of each latent signature as tissue-specific DMBs. I ran NMF on my methylation reference atlas for all tissue samples, filtered to only include the top 100,000 highest variance blocks, setting the number of underlying signatures, k to 22 (n healthy tissue types). Despite being completely unsupervised and not being provided with any tissue group labels, I found that NMF was able to deconvolute the tissue methylation reference into distinct latent methylation components

(LMCs), which generally correlated well with each individual tissue group (Figure 5.29. Furthermore, each LMC identified was comprised of fairly tissue-specific methylation signatures, as shown in Figure 5.30.

A similarly clear result was obtained for hmC, with clear tissue-specific markers identified from this unsupervised algorithm.

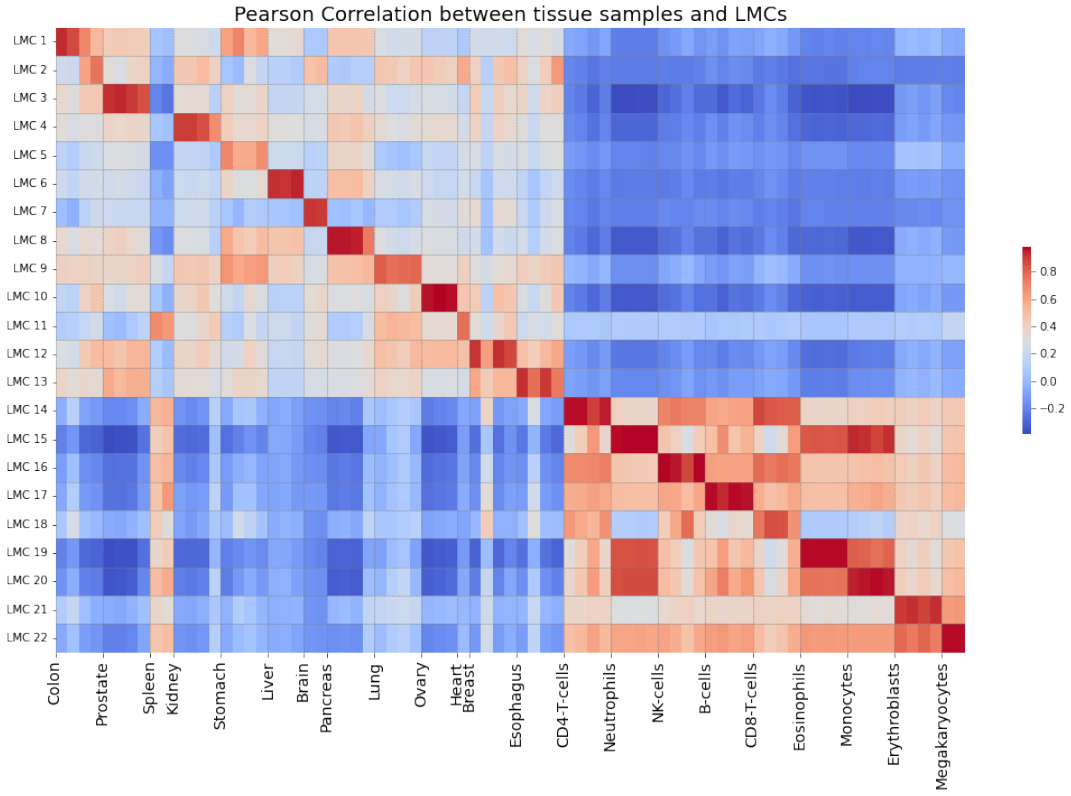


Figure 5.29: Correlation of each LMC with healthy tissue samples used to perform NMF

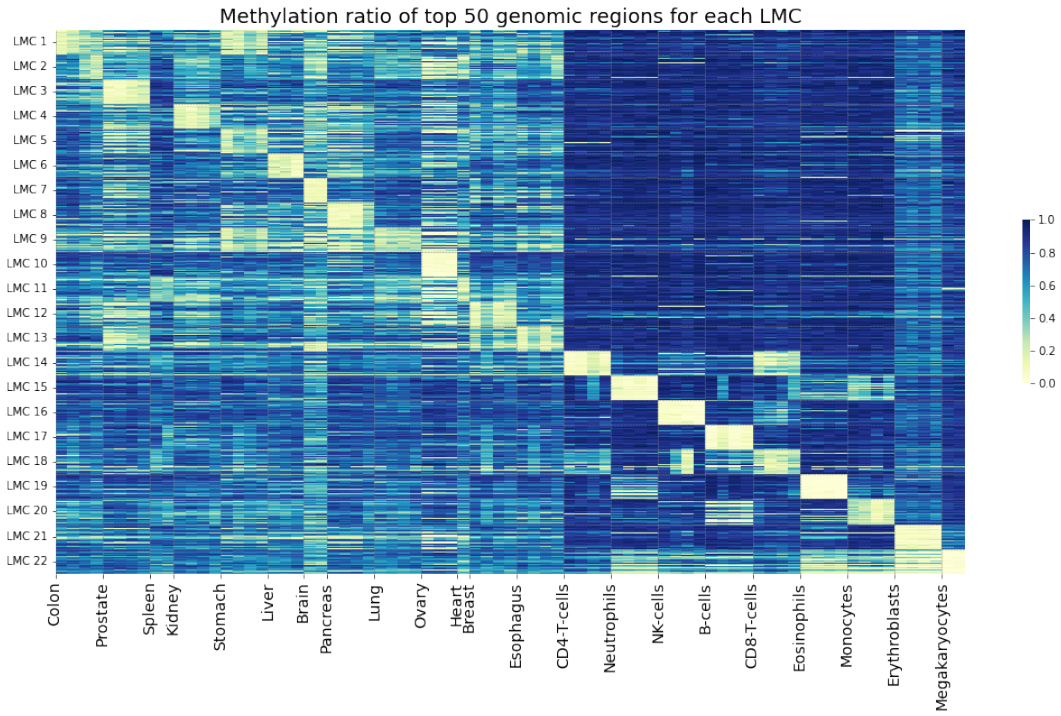


Figure 5.30: Methylation ratio for the top 50 highest weighted genomic features for each LMC in the above NMF.

5.6 Discussion

In this chapter, I introduce the collation of a valuable new public resource, the first of its kind to provide matched methylation and hydroxymethylation profiles across a wide variety of healthy and tumour tissues. The purpose of this dataset is two-fold: firstly to gain a better understanding of the interplay between mC and hmC, especially in relation to cancer development; secondly, to act as a comprehensive public source of tissue informative signatures for downstream tasks such as cfDNA deconvolution, and other models that can help distinguish cancer from healthy samples. The fact that when this work started no such references exist, yet there are now two other fairly comprehensive genome-wide methylation references published, highlights the general interest in such data.

By separately sequencing mC and hmC, we already see differences between these two epigenetic marks that would be missed using bulk sequencing technologies that cannot distinguish the two. For example, we expect from the literature to observe global hypomethylation in tumour samples, but we do not see this across all tissue types in this atlas. Instead, certain tissues such as kidney and brain exhibit an overall increase in tumour mC levels across all CpGs, with a counteracting decrease in hmC possibly masking this signal in bulk sequencing.

Brain tissue provides the clearest example of the distinct patterns of mC and hmC we can observe in such an atlas, with much higher global levels of hmC as has previously been reported in the literature. The variance of brain tissue hmC is also around an order of magnitude higher than for all other tissues, making it challenging to devise appropriately normalised methods that worked on brain and other tissues simultaneously. For example, the hmC segmentation could reasonably be biased by including brain; we may expect to see very different segmentation results after excluding it. I tested this and found this not to be the case, so reasoned the possible bias from including brain tissue is outweighed by the additional information it provides.

Segmentation and identification of differentially methylated blocks was also performed on methylation and hydroxymethylation datasets separately, allowing us to discover novel mC and hmC tissue-specific blocks with little overlap. Benchmarking and selecting the appropriate segmentation algorithm is challenging be-

cause we lack a ground truth: there is no single correct way to segment the human methylome, and we must instead derive reasonable metrics and heuristics that can help guide algorithm and parameter selection. There were extensive segmentation algorithms to choose from here, each with theoretical advantages but also practical limitations in terms of adapting them for genomic data.

hidden Markov model based segmentation has several key advantages over other methods: first and foremost it is quick, and computational cost scales well to large datasets, a necessary pre-requisite for any genome-wide method. This was true "out-of-the-box" with standard software package, and no special optimisation was required. Relatedly, it is modular and runs locally step-by-step, rather than attempting to find a global optimal solution, so is easily extensible. Secondly it was remarkably stable across repeated runs, and more robust than other algorithms to small parameter changes. Finally, HMMs are fairly interpretable and have been widely used in biological sequence data previously, such as for chromHMM segmentation of chromatin states [23]. More recently, a HMM-based DMR finding algorithm has also been published [167].

Tackling this segmentation problem required extensive visualisation of the spatial distribution of mC and hmC across the genome at the resolution of individual CpGs. It was interesting to observe the degree of coordinated methylation and hydroxymethylation across tissue types, with clear distinction of regions such as CpG islands surrounding genes. Rather than aggregating tissues by type, I chose to include all tissue biological replicates in the segmentation. Despite this increased variance, the segmentation algorithm still seemed to perform well (based on manually inspecting the inferred blocks) and consistently identified regions of coordinated methylation that are visibly distinct.

The fact that the identified DMBs are clearly tissue-specific and overlap with tissue-specific genes is itself a validation that this data compression has captured useful biological information. The enrichment of tissue-informative hypomethylated DMBs in enhancer regions and with tissue-specific histone modifications agrees well with another recent methylation atlas [91]. The majority of such regions fall outside of methylation arrays, which is further justification of this genome-wide approach.

Before construction of this atlas, hydroxymethylation as a tissue-informative marker was fairly under-characterised, and it was unclear what to expect when calling differentially methylated blocks on hmC data. Surprisingly, we saw fairly clear tissue-specific hyper DMBs on hmC for the majority of tissues, which are clearly enriched in tissue-specific genes. Hyper DMBs for hmC were enriched in enhancer regions and across gene bodies, which agrees well with results in a recent study of genome-wide hydroxymethylation in tissue samples [168]. The enrichment of hmC DMBs in H3K4me1 chromatin in a tissue-specific manner, also agrees well in both studies.

This was perhaps more surprising for tumour DMBs: a subset of tumour types such as kidney and ovary tumour showed faint hyper DMBs for hmC, despite the general observation that hmC is globally reduced in cancer. These hmC hyper DMBs were enriched with tumour-specific gene expression for a subset of cancer types. We lacked hydroxymethylation data during the analysis of cell-free DNA, but the tissue and tumour-informative regions identified here for hmC may be informative for future cancer diagnostic models. Now we have the technology to distinguish methylation and hydroxymethylation profiles through sequencing with TAPSbeta and CAPS separately, a future direction of this research will be to investigate whether these tissue and tumour-specific mC and hmC markers are observed in cell-free DNA.

I found no truly pan-cancer DMBs, regions that were differentially methylated in all tumour samples and no healthy samples. However, in performing this search I did identify several mC DMBs which were shared between more than half the tumour types. It is particularly interesting that the majority of these were hyper DMBs, despite hypomethylation showing much more clearly overall. Such regions could potentially be useful for cancer diagnostic tests in the blood, where it is easier to detect an increase in signal rather than an absence. Further investigation into these regions is underway, as is a study of which tumour groupings tend to share DMBs most frequently.

The DMB calling method used here was simple yet effective for our purposes. More generally however, the presence of this manual feature selection step before any downstream such as cfDNA deconvolution is problematic, due to its potential to introduce bias. I therefore investigated the range of automated features selection

methods that could be applied to DMB calling. These attempts were largely unsuccessful, and DMBs identified by these methods were not particulary tissue informative. The usual limitations most likely apply: a much larger dataset with more biological replicates per tissue may improve future iterations. It does seem plausible that replacing this manual feature selection step with an end-to-end model that can itself select informative features, may be a useful improvement in this domain.

A different form of unsupervised feature selection which worked surprisingly well for both mC and hmC was non-negative matrix factorisation. Despite being provided all tissue samples and no tissue labels, this algorithm identified latent methylation components that clearly corresponded to the atlas tissues. NMF is distinct from other DMB calling algorithms used as it simply extracts a weighting for every genomic position, effectively producing a tissue signature rather than a finite list of DMBs. By taking a manual cutoff of the top weighted regions, we can clearly extract tissue-specific DMBs using NMF.

In selecting tissues for such a reference, there is inevitably a trade-off between depth and breadth of atlas; we chose to prioritise more comprehensive coverage of body organ systems at the tissue level at the expense of decreased cell-type resolution. An additional use of NMF for this tissue atlas could be to separate current tissues into their constituent cell-types. By running NMF on each tissue individually, the resulting LMCs may correspond to methylation signatures of underlying cell types that make up this mixture, and help to remove blood cell contamination and other impurities.

Chapter 6

Predicting gene expression using genome-wide patterns of cytosine methylation and hydroxymethylation

6.1 Abstract

Methylation and hydroxymethylation are distinct, tissue-specific epigenetic marks whose deposition across the genome is known to regulate downstream gene expression. However, many details of this regulation remain unclear, such as how this linkage changes as a function of location along the genome relative to genes, and the tissue specificity of this relationship. We also do not know how much this linkage differs between hydroxymethylation and methylation, and how distinct a role hydroxymethylation plays in regulating gene expression. Having constructed a comprehensive tissue reference atlas in the previous chapter and used it to demonstrate that mC and hmC signatures are distinct, in this chapter I use this atlas to probe the functional relationship between genome-wide mC and hmC signatures and gene expression. I first examine the correlation of these epigenetic marks with gene expression, as measured by RNA-seq transcript abundance, and find that mC and hmC gene body show strongest correlation with gene expression, albeit in opposite directions. Hydroxymethylation over gene body demonstrates the strongest tissue-specific effect. I then construct a variety of models to quantify this relation-

ship, and investigate how well we can predict gene expression information from these two epigenetic marks alone. I find that a deep learning classifier based on window-aggregated mC and hmC features exhibits best performance. This model can predict cancer gene expression in new held out datasets with reasonable accuracy, and reinforces hydroxymethylation of gene body as the most predictively useful feature.

6.2 Introduction

DNA methylation and hydroxymethylation are important epigenetic regulators of gene expression. A hepatocyte cell in the liver and a neuron in the brain have an identical genetic sequence (barring somatic mutations), yet these cells perform highly distinct and specialised roles. Although it is assumed that methylation signatures orchestrate highly tissue-specific gene expression, little is known about the mechanism or regulatory syntax of this interaction. Because methylation changes are early hallmarks of cancer, we need a better understanding of how changes to epigenetic patterns alter downstream gene expression, before we can predict the effects of any methylation changes detected. There is significant interest in developing cancer liquid biopsy tests that focus on methylation changes in cell-free DNA [58]. Improving our understanding of the mapping between upstream epigenetic changes and downstream disease-specific changes to gene expression can help us mechanistically interpret and infer the effects of epigenetics-based diagnostic assays.

Many gene regulatory steps lie between the genetic and epigenetic information encoded in DNA in the nucleus, and the transcription of this information into messenger RNA for translation into proteins. The exact influence and regulatory role of epigenetics upon this process is unclear, and the associations have not been quantified. Many questions remain unanswered, such as: where in the genome is DNA methylated to alter tissue-specific gene expression, and does this positional information vary across genes? Is this spatial distribution of methylation tissue-specific or constant across tissues? How large a role are these epigenetic changes playing in altering downstream expression? Furthermore, the role of hydroxymethylation in gene expression regulation is relatively unknown: does this

epigenetic mark have a distinct influence, and how do these two chemical modifications interact to drive tissue-specific gene expression?

To begin to shed light on this process, I explore the relationship between genome-wide mC and hmC from tissue reference with gene expression, first by correlation analysis then through a variety of predictive models. Having collated a wealth of high-quality data in our tissue mC and hmC reference atlas, we can use this resource to probe the downstream consequences of the observed patterns in the cell, and gene expression is the central determinant of cell phenotypes.

The process of transcription of accessible DNA on chromatin into messenger RNA (mRNA), and the subsequent translation of mRNA into protein activators is a dynamic and heavily regulated process in each cell. There exist many layers of regulation and feedback loops that act at different stages to fine-tune protein production to suit environmental requirements. In reality, gene expression exhibits large temporal and spatial variation, and any attempt to measure gene expression is necessarily an imperfect snapshot at an artificially fixed time point. Indeed, even the term gene expression is a convenient umbrella term that actually encompasses a wide range of distinct processes related to gene activity. Several studies have attempted to predict gene expression previously, but use different measurements for gene expression which occur at different stages along the pathway from DNA to protein translation. I use the gene expression proxy of RNA-seq transcript abundance, which is a widely used gene expression measure as it is fairly late along the pathway, and represents mRNA abundance at the step before protein translation begins.

I present here a model mapping CpG level methylation and hydroxymethylation to RNA-seq transcript abundance. Specifically, I first examine correlation between mC and hmC and gene expression separately, and how this interaction changes across functional regions of the genome. I then attempt to better quantify the observed relationship by building predictive models of RNA-seq transcript abundance using mC and hmC.

Despite the many layers of regulation that lie between nuclear DNA and RNA expression, several recent papers have demonstrated that the gene expression program of a tissue can be predicted with reasonable accuracy from DNA sequence

alone. In 2021, Avsec *et al* [104] demonstrated that a deep learning (Transformer-based) model with a wide receptive field can predict RNA-seq as measured by Cap Analysis of Gene Expression (CAGE) from DNA sequence alone, with a Pearson correlation up to 0.85. A recent paper by Linder *et al* achieves a similar model correlation score when predicting RNA-sequencing coverage across the genome [112], in the latest evolution of gene regulatory models from Basset to Basenji to Borzoi. Because DNA sequence is identical in each tissue and cell type, these sequence based models are necessarily multi-task models, making multiple predictions for distinct tissues simultaneously. Whilst they may achieve overall high model accuracy averaged across all genes, these models are not tissue-specific, and it is possible that some tissue-specific effects are lost. Epigenetic aberrations that occur early in cancer in a tissue-specific manner may well influence the gene expression program in that tissue to encourage unchecked proliferation, but the effects of these tissue-specific changes would be missed by such a generalist model.

I focus instead on examining the relationship between epigenetic marks, which are inherently tissue-specific, and downstream gene expression programs. Fewer predictive models examining this specific relationship exist. The two closest comparisons are CHALM [169], and a sub-section of the hydroxymethylation based hmC-SEAL study [168]. The CHALM study devises an alternative methylation summarisation method and demonstrates this is a better predictor of gene expression, however they do not demonstrate that these predictions are tissue-specific. By contrast, the hmC-SEAL predictive model does demonstrate tissue-specificity, but is based on aggregated hmC over gene-body alone. In this chapter I develop a gene expression prediction model with improved resolution and performance, by devising a novel encoding that combines mC and hmC at 1 kilobase resolution. I demonstrate the model has consistently good performance on held-out tissues, and can generalise to new datasets by predicting cancer gene expression.

6.3 Obtaining Linked Epigenetic and Gene Expression Data

We have seen that that methylation and hydroxymethylation levels vary widely across different functional regions of the genome. I first sought to better charac-

terise how the relationship between these two epigenetic marks and gene expression changes across different functional genomic regions. I downloaded public RNA-seq transcript abundance data for all genes from the Human Protein Atlas (HPA) platform [170], which is consensus transcriptomics data from HPA and GTEx [163, 171]. This data is normalised using log transcripts per million (logTPM) [172]. Linking functional regions to these genes is not straightforward, and requires some assumptions to be made. Cis-acting features such as promoters can be fairly confidently assigned to their nearest gene, but elements such as enhancers can influence gene expression from a much greater distance, and in a tissue specific manner. Indeed, the assignment of enhancers to their target genes is in itself an active field of research [173, 174, 175]. I tried several methods of linking these distal elements to genes, including bespoke databases linking CpG Island shores and enhancers to genes, such as GeneHancer [176]. However, only a subset of all 20,162 genes contain a high-confidence linkage in these datasets. When trying to capture linkages of multiple elements including enhancers, CpG islands, promoters and transcription factor binding sites, we are left with fewer than 10,000 genes.

To increase the number of covered genes, I subsequently took a simpler approach: taking the nearest proximal gene to each consensus enhancer and promoter identified in a recent ChromHMM update, as well as the gene body. Gene names, start and end positions were taken from the recent consensus database of all Matched Annotation from NCBI and EMBL-EBI (MANE) genes [177]. After filtering by MANE genes covered by all these elements, and a containing high coverage CpG methylation and hydroxymethylation information, we end up with 15,894 genes. The distribution of mC and hmC across these three genomic elements is quite distinct, and varies across different tissue types, as shown in Figure 6.1. Promoter methylation and hydroxymethylation levels are consistently low, whilst enhancer methylation and hydroxymethylation show opposite trends. We also see wider variation in methylation across enhancers and gene bodies, whereas hmC tends to be quite consistent. Brain displays a distinct distribution of mC and hmC, and blood cell types and tumours are quite distinct from healthy solid tissues.

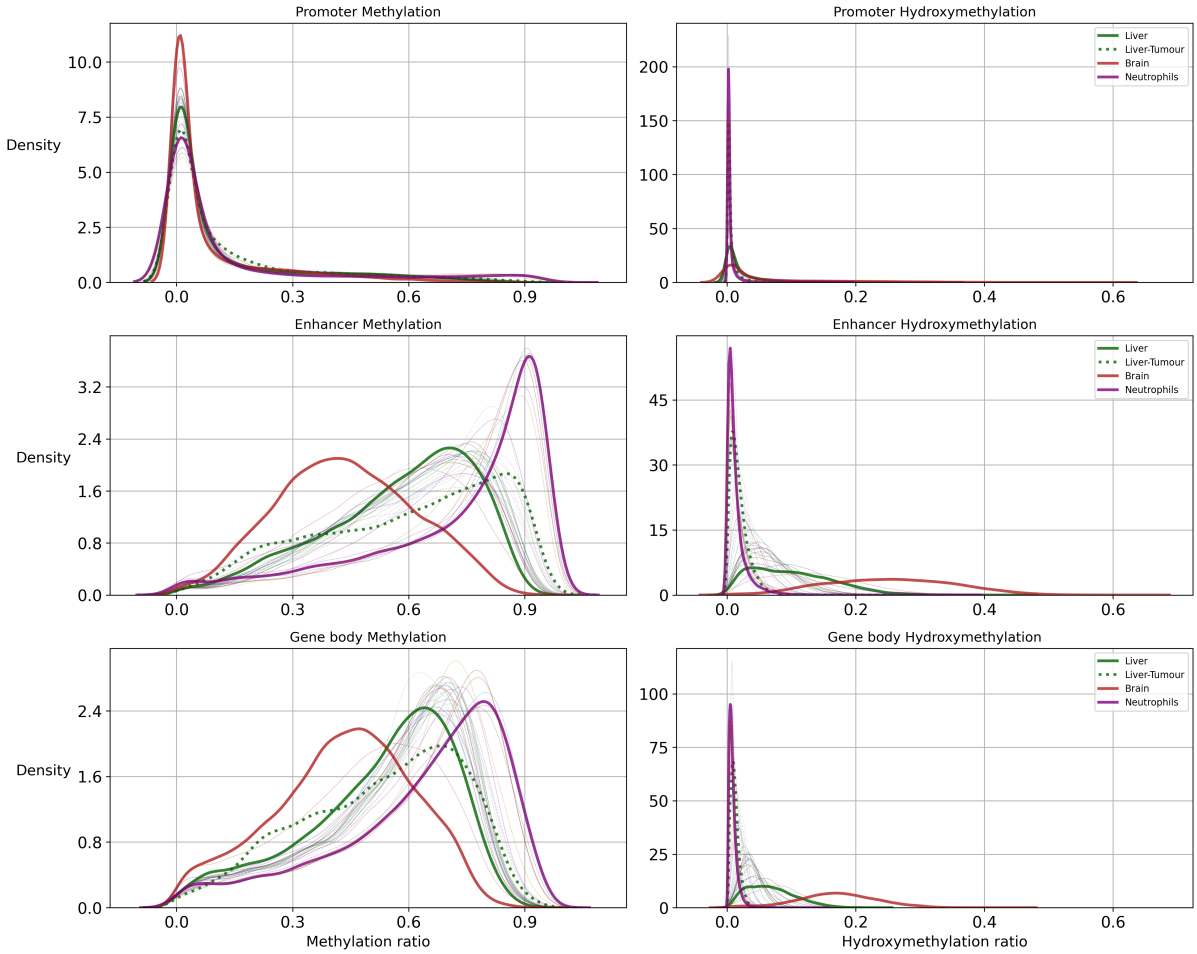


Figure 6.1: Probability distribution of methylation and hydroxymethylation across promoters, enhancers and gene bodies of the 15,894 covered genes in this dataset, calculated using kernel density estimation. For clarity, four representative tissues/cell types are highlighted to show the distinct distributions of solid tissues, brain, tumour types, and blood cell types. Remaining tissues are shown in grey.

6.4 Correlation of methylation and hydroxymethylation with RNA-seq transcript abundance

I examined correlation of mC and hmC with RNA-seq transcript abundance across these three genomic elements. This correlation is illustrated for a single tissue (lung) in Figure 6.2, and is summarised across all tissues in Figure 6.3. Enhancer, promoter and gene body methylation was negatively correlated with RNA-seq transcript abundance across all tissues, but to varying degrees. Promoter hydroxymethylation was also negatively correlated, whilst enhancer and gene body hydroxymethylation were positively correlated with RNA-seq transcript abundance.

To further probe this relationship, I separated genes into tissue-specific and "housekeeping" genes. Tissue specific genes are defined on the HPA platform as genes with "elevated expression in a particular tissue compared to other tissues" [171]. A total of 10,930 genes match the HPA criteria of "at least four-fold higher mRNA level in a particular tissue compared to the average level in all other tissues", whilst a remaining 8,194 genes show low tissue-specificity, and are here termed "housekeeping" genes. As seen in Figure 6.3, we observe distinct trends for housekeeping and tissue-specific genes: the genomic feature showing largest (negative) correlation across all housekeeping genes is gene body methylation, whilst gene body hydroxymethylation shows the largest (positive) correlation with tissue-specific genes. There is an important distinction to be made between a feature having high magnitude correlation versus highly tissue-specific correlation. Gene body methylation was generically high across all tissues, whereas gene body hydroxymethylation showed both high and tissue specific correlation with RNA-seq transcript abundance.

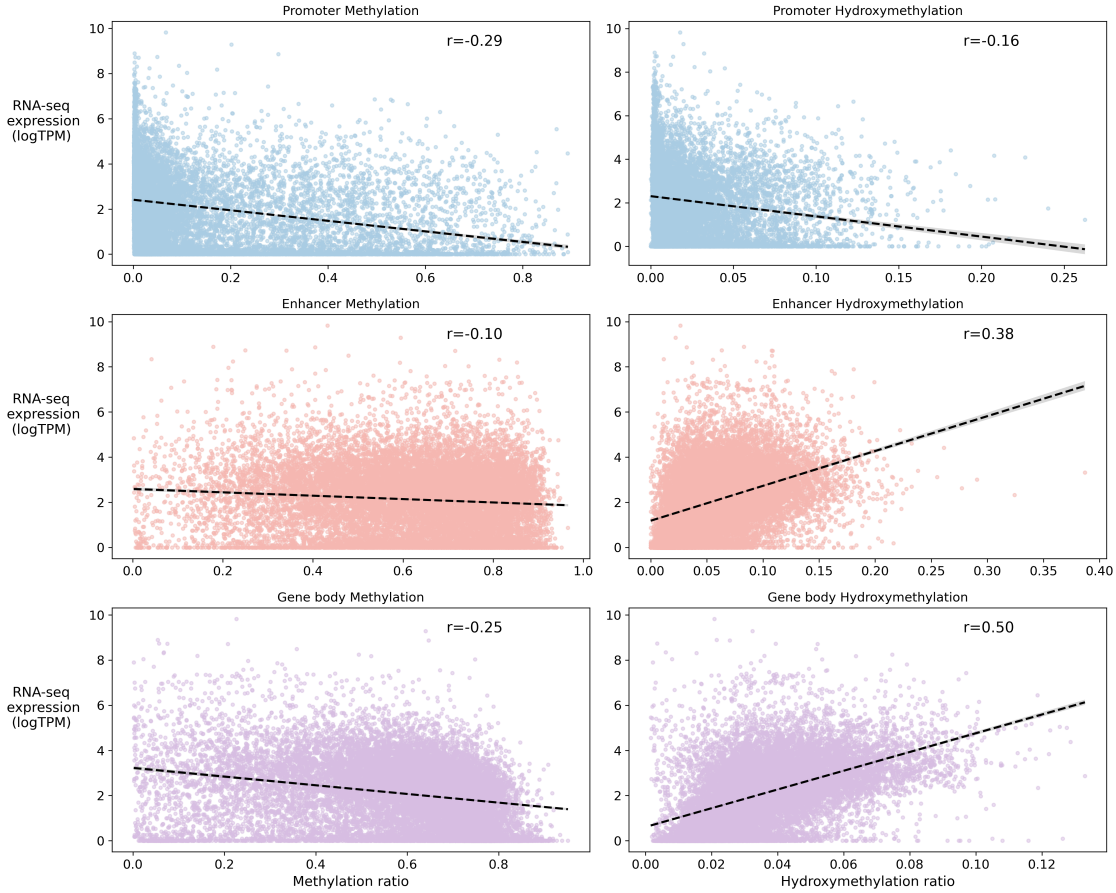


Figure 6.2: Pearson correlation of methylation and hydroxymethylation across promoter, enhancer and gene body genomic elements of all 15,894 genes. This figure illustrates correlation for a single tissue (lung), which is a representative example of the general trend in correlation across all tissues in the tissue atlas. A linear line of best fit and Pearson correlation value is also shown.

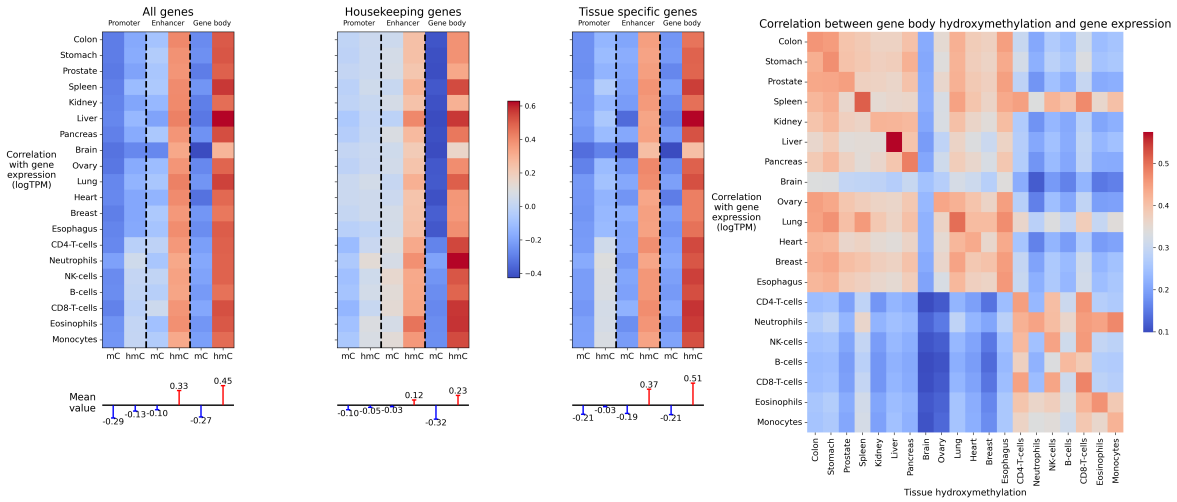


Figure 6.3: Pearson correlation of methylation and hydroxymethylation across promoter, enhancer and gene body genomic elements of all, separated by tissues. The left-hand plot shows this correlation for all 15,894 genes in our filtered dataset, then moving right we see the same correlations for the subset of housekeeping and tissue specific genes. Beneath each plot is the mean correlation value for each feature across all tissues. On the right hand side, the cross-correlation of gene body hydroxymethylation of every tissue in the tissue atlas with every RNA-seq tissue expression profile, thus demonstrating the tissue specificity of this epigenetic mark.

6.5 Predictive models of gene expression

To further explore this relationship, I built predictive models of RNA-seq transcript abundance using mC and hmC as predictive features. To obtain sufficient data for model training, I used genomic features from all genes across all tissues in our dataset, thus creating a dataset of size m tissues $\times$ k genes. Models were trained using leave-one-tissue-out cross-validation, whereby all genes from a single tissue are held-out for model evaluation, and we train on all k genes across all $m-1$ tissues. I tried various 'shallow' machine learning classifiers such as Elastic-Net regression, gradient boosting and random forests, evaluating models using R^2 and mean absolute error (MAE) across different feature combinations. I initially included CpG Islands and CGI shores as additional features, but the the additional information provided by these features was outweighed by the decreased

training set size resulting from the reduction of covered genes. Ultimately the best performing models were random forest models trained on mC and hmC enhancer, promoter and gene body, which displayed an R^2 range of 0.55 to 0.3 across held-out tissues. The model feature importance closely matched correlation statistics, with highest feature importance assigned to hydroxymethylation of gene body. Models trained on tissue-specific and housekeeping genes separately performed slightly worse overall, but also corroborated the feature importance of hmC gene body for tissue-specific gene models, and mC gene body for models built on housekeeping genes.

To further pinpoint the locations where methylation and hydroxymethylation contribute to gene expression prediction, I replaced functional genomic elements with 1-kilobase bins surrounding each covered gene. This had the additional advantage of increasing the size of the training set, thus making this problem amenable to more complex models. I used MLML [161] to calibrate these ratios and additionally obtain the remaining unmethylated fraction (umC) in each 1kb bin, which I include as an additional feature with the same binning.

To maximally capture interplay between mC, hmC and umC in influencing gene expression, I used an encoding that combined all three epigenetic marks into a single stacked array with 3 rows and i genomic windows. Taking inspiration from the Enformer and Borzoi gene expression prediction models [104, 112], I additionally widened the receptive field, considering epigenetic signatures 20kb either side of each gene body. A single training data point using method 1 encoding is illustrated in Figure 6.4. .

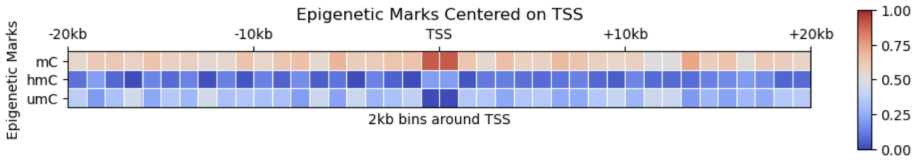


Figure 6.4: A single example of an input data array for a random gene and tissue, to illustrate the input data for our ConvNet model.

6.5.1 Model architecture and training

To capture spatial interactions between these features, I used a Convolutional Neural Network architecture with 4 convolutional layers, each separated by ReLU activation to add non-linearity and max pooling layers. The high-level features extracted by these convolutional blocks are then flattened and passed through 3 fully connected layers, reducing the output dimensionality to one. Residual connections are added after each convolutional layer to improve gradient flow through the network with shortcut connections that skip one or more layers, as is commonly used in architectures such as ResNet50 [178]. PyTorch (version 1.12.1) is used for all models.

The model objective is to minimise mean squared error (MSE) between observed and predicted logTPM, and is trained using an Adam optimizer with a learning rate scheduler to adapt the learning rate to the loss landscape. The model is first evaluated using using leave-one-tissue-out cross validation: for each tissue t, I held out all genes for tissue t and trained on all genes for remaining $T-1$ tissues. After this evaluation, a global model is trained on all tissues and all genes, which is then evaluated on tumour data.

I tried two distinct feature encoding methods:

1. Mean mC and hmC 10 or 20 1-kilobase bins either side of the gene transcription start site (TSS).

2. Dividing genes into 10 or 20 equal length windows, such that each window has length $\frac{\text{gene_length}}{20}$, and calculating mean mC and hmC in each window, as well as 20kb upstream and downstream of each gene TSS in 1 or 2kb bins.

The format of the input data for encoding method 2 is illustrated for a randomly chosen gene in Figure 6.5. Gene body lengths vary across several orders of magnitude, so the length of each gene body feature in method 1 will vary as a function of gene length. A feature 2kb downstream of TSS may correspond to the first exon of gene A, whilst corresponding to the end of gene B, and we expect different mC and hmC environments for these two scenarios. Method 2 attempts to avoid this misalignment by dividing the gene into equal windows, such that the first window always corresponds to the first section of each gene, thus maintaining

feature correspondence across all genes. Both methods performed similarly, but the model feature importance of method 2 is easier to interpret, so I focus on method 2 here.

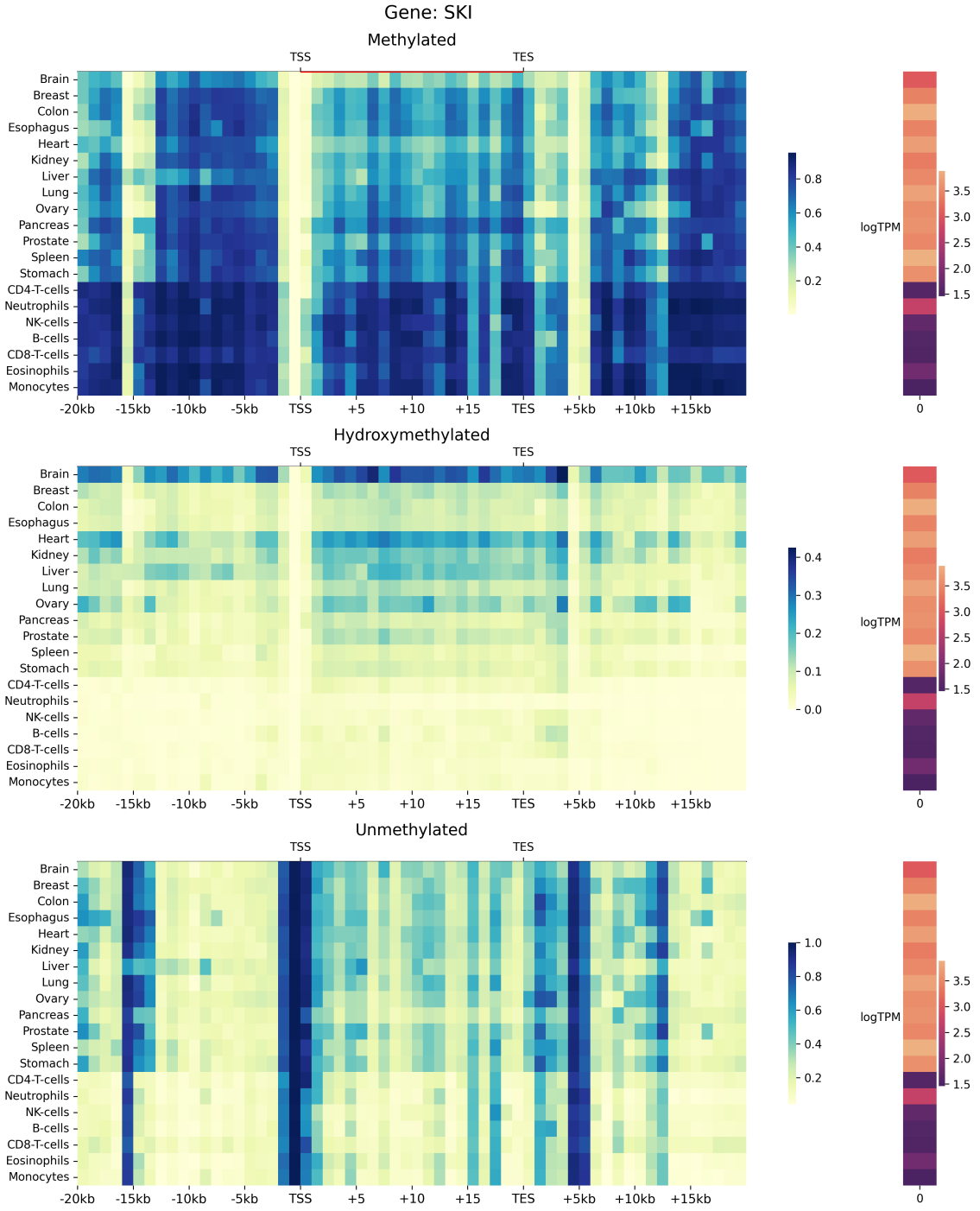


Figure 6.5: Methylation, hydroxymethylation and unmethylated fractions surrounding the TSS of a randomly selected gene (SKI), binned into 1kb bins for 20kb upstream and downstream of gene body. The gene body is divided into 20 equal length segments. This is used as input data to a predictive model, where the target is RNA-seq Gene expression (logTPM) as shown on the right.

The resulting models demonstrated improved performance on each held out tissue, with R^2 scores varying between 0.79 to 0.60 for different tissues. Model predictions with all three epigenetic marks combined within a single array were significantly more accurate than models trained on any single epigenetic mark alone, with an average R^2 difference of 0.1. Model predictions with a widened receptive field up to 40kb showed an average R^2 increase of 0.05. The most stable and accurate model used mC, hmC and umC binned into 1kb bins 20-kilobase either side of gene body, and each gene divided into 20 equal windows. Results from this model are illustrated in Figure 6.6. In addition to this model training scheme, I additionally reframed this as a classification problem by binning RNA-seq logTPM abundance into three equal bins. The resulting three-class classifier achieved comparable performance to the regression based models mentioned here, and may be revisited in subsequent iterations.

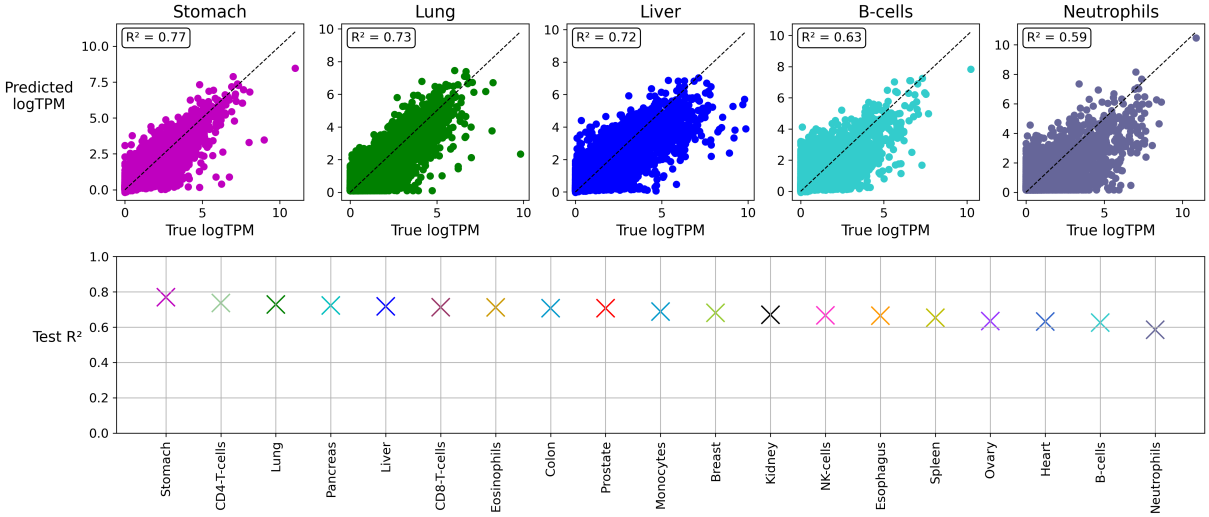


Figure 6.6: Model performance of gene expression prediction model, evaluated on all genes for a single held-out tissue. R-squared plots for best and worst performing models are shown on the top panel. The lower panel shows the distribution of R-squared values.

Notably, models trained to predict held-out brain tissue had considerably worse performance than all other tissues, so brain was removed from the training data of final models. Brain cell types exhibit a distinct interplay of hydroxymethylation and methylation, so it is possible that the relationship between methylation and hydroxymethylation and its' effect on downstream gene expression is fundamentally different in brain compared to other tissues.

6.6 Feature importance of gene expression prediction model

It is not straightforward to obtain featuring importance for deep learning models, but one common approach is to use Shapley Additive Explanations values [179]. This method uses a game-theoretic approach to weight the importance of each input feature in the resulting model. I used SHAPley feature importance of the best performing model for both 1kb and 2kb-binned input data demonstrated a similar spatial distribution, as illustrated in Figure 6.7. The features with highest importance are hydroxymethylation gene body, and this feature importance is distributed across the gene body in hmC. By contrast, the most important mC feature is right at the start of the gene body, closest to the gene TSS.

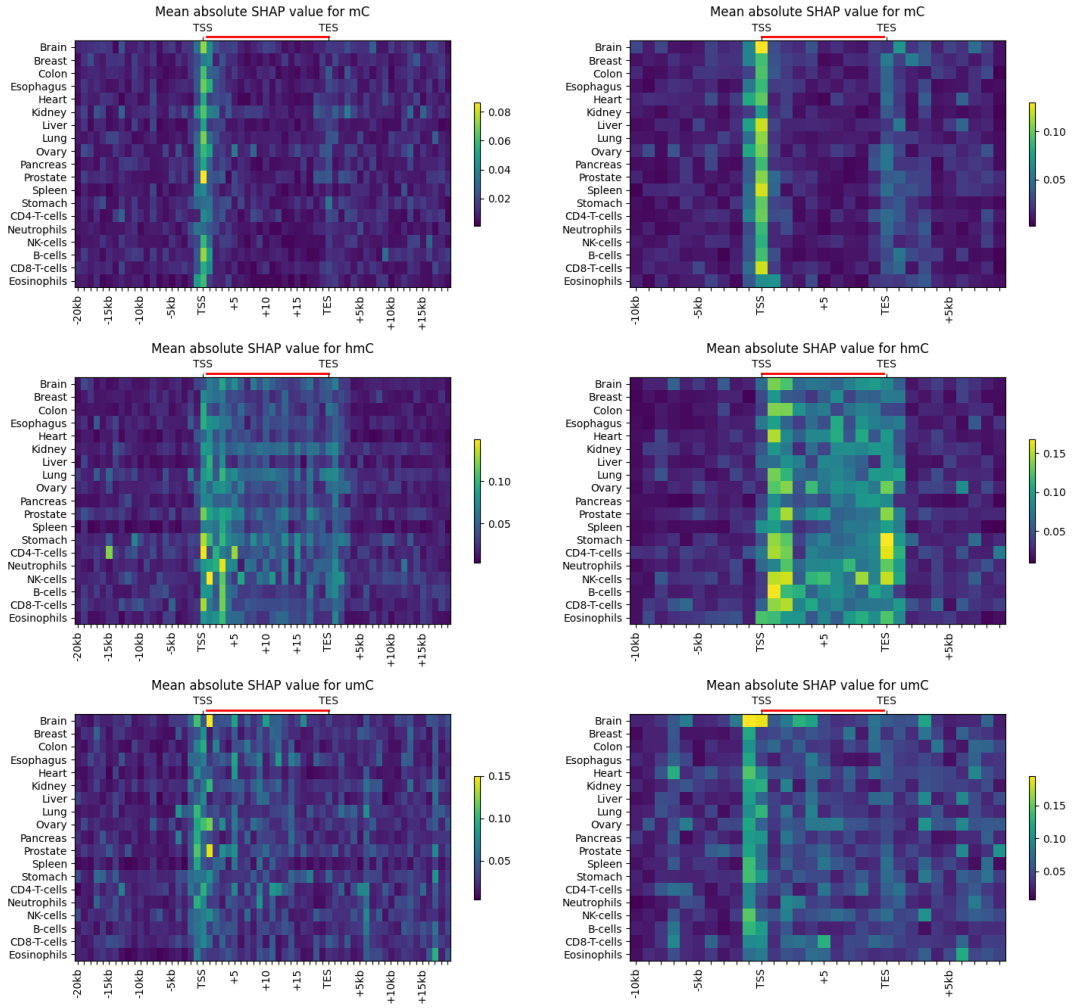


Figure 6.7: SHAPley feature importance for mC, hmC and umC across all tissues for the 1kb model (left) and the 2kb model (right).

6.7 Validating predictive model on new cancer gene expression data

To further validate this trained model, I sought to predict gene expression in two entirely new datasets of cancer gene expression. For each of the ten tumour types in our tissue atlas, I downloaded RNA-transcript abundance data for the closest matching cancer type from two distinct TCGA datasets. The cancer gene expression dataset first was taken from the UCSC TOIL platform which recomputes disparate RNA-seq data in a unified manner [180], reporting normalised transcript abundance using logTPM; the second was taken from HPA platform and reports transcript abundance using FPKM (fragments per kilobase of transcript per million reads mapped). I used a model trained on all healthy tissues to predict the RNA-seq gene expression of each tumour type, using mC, hmC and umC alone as predictive features. To avoid overly-inflated scores from housekeeping genes whose expression varies little between tissues, I filtered to only include the 10,930 tissue-specific genes. The resulting predictions were heterogeneous across different cancer types, but in general the model predictions for each cancer type showed high and tumour-specific correlation with the true transcript abundance, as shown in Figure 6.8. Our trained model was particularly good at predicting prostate, kidney, liver and brain tumour expression levels, and poor at predicting stomach, ovary and breast tumour expression levels.

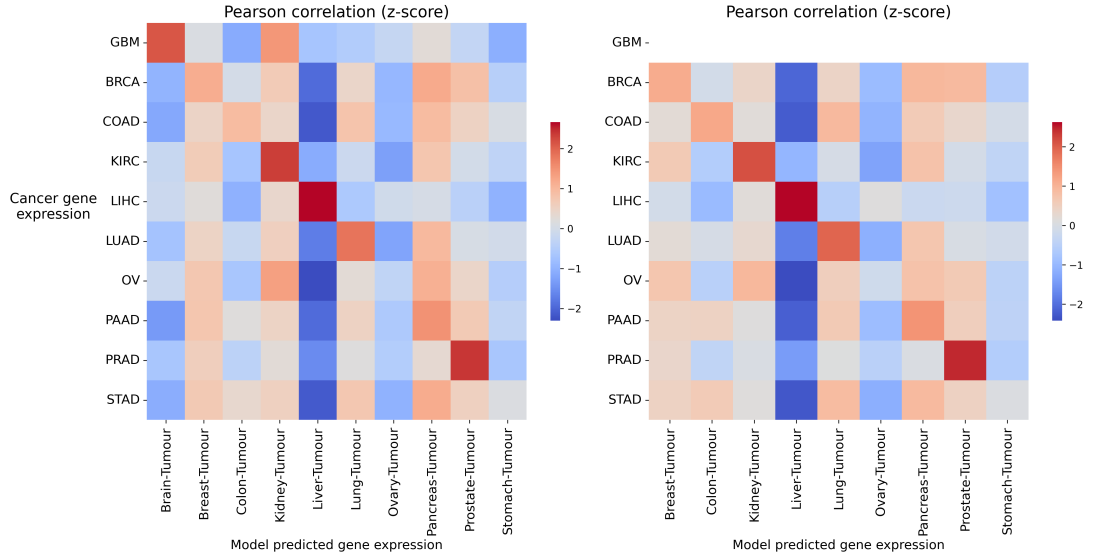


Figure 6.8: Z-score of Pearson correlation for model predictions with two independent cancer gene expression datasets. GBM was missing for the second dataset. For each tumour type in our reference, I identify the best matching cancer type with public gene expression data, for example brain tumour and glioblastoma (GBM). I then predict gene expression for all genes using mC and hmC distribution in brain-tumour and correlate these predictions to observed GBM gene expression. To benchmark this score and investigate its tumour specificity, I also compare gene expression model predictions to all other cancer types.

6.8 Discussion

This chapter presents the first investigation of the downstream consequences of the methylation and hydroxymethylation signatures observed in our tissue atlas. The trends identified here generally agree with the literature consensus in terms of correlation of mC and hmC with gene expression across different genomic elements. Methylation of promoter and enhancer elements was negatively correlated with gene expression across all tissues, concomitant with its established role as a gene silencing mark. Although the original consensus was that gene body methylation is also repressive of gene transcription [9, 181], recent evidence to partially contradict that, suggesting that at some CpG Islands within gene bodies methylation is actually associated with increased transcription [21]. Nevertheless, we observe a negative correlation between gene body methylation and gene expression across all tissues also negatively correlated with gene expression. A possible explanation is this correlation analysis used a single mean aggregated methylation value to represent each feature (for each gene and tissue combination), so lacks the spatial resolution to distinguish a local positive effect due to CGIs at specific loci within gene body, from the overall negative correlation. A clear next step is to repeat this correlation analysis at greater spatial resolution using the 1-kilobase bins I defined.

This study is the first attempt to use a deep learning model to predict RNA-seq transcript abundance using epigenetic data alone. There was no particular expectation how well such a model should perform, or indeed how strongly linked these two cellular processes are. Many fine-tuned gene regulatory processes lie between chemical modification of DNA with a methyl or hydroxymethyl group, and the transcription of genes into mRNA. It is therefore interesting to observe that RNA-seq transcript abundance in a tissue unseen by the model can be predicted with reasonable accuracy using the methylation and hydroxymethylation patterns in that tissue alone. It is informative that finding an encoding that incorporated spatial interactions between epigenetic marks through convolution operators proved important for model performance. Similarly, it is notable that widening the receptive field to 40kb either side of the gene TSS model prediction accuracy, as was found by [104] and [112].

Of particular note is the importance of hydroxymethylation, whose role in gene regulation is still largely unexplored, in both correlation analysis and subsequent predictive models. I showed previously that the majority of tissue-specific DMBs for hmC are distributed evenly across the gene body, in contrast to mC which are distributed more sharply around the TSS. Interestingly, we observe a similar distribution for feature importance of these gene expression prediction models around the gene body. Hydroxymethylation features have highest feature importance, but this is distributed fairly evenly across the gene body. This result supports the evidence for positional differences between mC and hmC, and also corroborates the distinct role that hydroxymethylation plays in marking the genome and regulating downstream gene expression. The results obtained here agree well with a recent large study on tissue hydroxymethylation, that lacked CpG resolution data so focused on functional regions [168]. They find that hmC is depleted at TSS and enriched across gene bodies, and that gene body hmC correlates well with gene expression in a tissue-specific manner.

Further validation of this model on new, independent gene expression data is clearly needed to fully evaluate its usefulness. More validation data will allow us to better quantify model performance on different subsets of genes, assess model uncertainty, and clarify how useful such a model could be for downstream tasks. Future work will also assess how model performance changes when predicting gene expression using just the mC and hmC tissue-specific DMBs identified in the previous chapter. Perturbation experiments will be carried out to evaluate the model dependence on methylation hydroxymethylation at different genomic loci. These perturbation experiments explore the relationship between data input and predicted output as learnt by the model, by simulating a change such as a DNA mutation in the case of sequence based models, and modelling the predicted outcome on gene expression [104]. By performing similar perturbation experiments for this mC and hmC based model, we can begin to predict the effect on gene expression of epigenetic aberrations that often arise in early in cancer development [29].

More generally, as methylation and hydroxymethylation data becomes more abundant, a model that can reliably predict gene expression from epigenetic data

could be applied to cell-free DNA and aid non-invasive cancer diagnostics. Provided the mC and hmC profile observed in cfDNA could be deconvoluted into tissue-specific signatures, we may be able to predict changes in gene-expression programs related to cancer development.

Chapter 7

Discussion

The ultimate aim of this thesis is to improve computational methods and data availability to expedite cancer diagnosis. I take two complementary paths towards this end, firstly on applied diagnostics data, by improving methods for analysing clinical cfDNA datasets, with improved methods for deconvolution and multimodal classification. Secondly, with a deeper examination of mC and hmC biology and how it changes in cancer, and the construction of a new tissue mC and hmC reference atlas to aid this study. This will serve as a valuable resource that allows us to ask more detailed questions about these two under-characterised epigenetic marks.

Computational data analysis methods in biology are limited by and thus always lag experimental methods available to collect the data in the first place. We are at a time of rapidly advancing technical methods in genomic medicine, with high-coverage genetic and epigenetic sequencing becoming widely available and much cheaper. The breadth and resolution of these tools is improving, and we should develop computational tools that make full use of this improved resolution. For example, the development of TAPS sequencing of cfDNA allowed us to obtain higher-quality methylation calls across the genome, as well as genomic sequence and fragment length information simultaneously from the same sample. Many of the computational methods developed in chapters 3 and 4 of this thesis are motivated by the availability of this new, high quality data in the TAPS pilot study. The reduced degradation of this sequencing reaction meant that cfDNA fragments were better preserved, allowing the development of cfDNA classifiers where each

data point is a single read. This is in contrast to read pile-up approaches used previously, which were necessary for low-confidence methylation calls resulting from low-coverage bisulfite sequencing.

Similarly, the possibility of obtaining genome-wide mC and hmC tissue signatures at CpG-resolution motivated the development of a new segmentation algorithm that can better define the functional building blocks of methylation and hydroxymethylation. Computational methods should stay as close to the underlying biological molecules as possible, and minimise introduction of artefacts and bias that can obscure this view. We can think of models as essentially learning a mapping from biomolecule to phenotype of interest, and we want to minimise the number of steps between input and output, or obstacles that can obscure the view.

I first develop an analytical pipeline to extend cell-free DNA deconvolution on genome-wide methylation data, taking a more unbiased approach to previous methods that focus on pre-defined loci on methylation arrays. A necessary prerequisite for this process was to collate a tissue methylation reference atlas from public data. At the time no such reference for genome-wide methylation data existed, and this atlas proved a valuable source of tissue-informative markers. Many of the tissue-informative markers I subsequently used for deconvolution were not present on standard methylation arrays used previously. These genome-wide tissue informative markers allowed me to deconvolute cfDNA and obtain results that agreed with consensus literature, and separate liver cancer cfDNA samples from non-cancer controls.

I evaluated the available reference-based tissue deconvolution algorithms, and showed the two main methods of NNLS and CelFiE largely agree, before devising a new method that utilises the power of supervised learning. The clearest limitation of the results I obtained is the small sample size and the absence of testing on new data. Although the results are promising, we must be hesitant to draw conclusions about utility without evaluating the generalisability of these methods to new data. Without new data, I was unable to demonstrate a marked improvement in distinction of cancer from controls using this newly devised method compared to previous methods: the predicted tissue contribution for HCC agreed with previous results, and allowed us to distinguish HCC cfDNA samples from control samples.

The general agreement of these three methods, across multiple references indicates that the inferred contribution may indeed be a true representation of the underlying biology. Another deep learning approach to tissue deconvolution was published around the same time, independently of my work. Their approach is very similar to mine, and our findings agree well. This is helpful as an independent validation of my method.

Similarly, no deconvolution method was able to detect a PDAC-specific increase in pancreas contribution, making distinction of PDAC from control more difficult. This was helped somewhat by consistent (but not statistically significant) changes in immune cell contributions being inferred for PDAC. Again, the consistency of this negative result may indicate that there is indeed very little pancreas-specific signal in the observed cfDNA mixture. There are many plausible biological explanations for this. Pancreas is a smaller tissue with less direct access to the bloodstream than liver, so we may expect a lower rate of cfDNA shedding. Indeed, several studies have successfully detected liver cancer from cfDNA [82, 83, 124], but relatively few, if any, studies have reliably detected pancreatic cancer. Pancreatic cancer remains one of the most deadly cancers, primarily due to going undetected until late stage disease, so further method development is desperately needed here. Intriguingly, distinct hydroxymethylation signatures have been detected for early stage PDAC [182]. In the analysis of our novel tissue reference atlas we identified mC and hmC markers for pancreas and pancreas-tumour, and replicating this finding is a priority among the future directions.

Cell-free DNA deconvolution is fundamentally an exploratory approach. The true tissue contribution to the cfDNA pool is next to impossible to obtain for real patient data, so we must assume that the assigned contribution is informative. It is therefore particularly important, before applying cfDNA deconvolution on real patient data, to ensure that the deconvolution method is working as expected, and minimise the introduction of bias or data artefacts. However, this lack of validation data makes method optimisation particularly tricky for this problem, as we lack feedback. The same was true for devising a novel segmentation algorithm for mC and hmC, and the subsequent DMB calling algorithm. A consistent theme in the problems tackled in this thesis is we are often "searching in the dark": attempting to find an optimal method or optimal parameters for a method where there is no

ground truth correct solution to optimise for. We must therefore be creative in devising ways to obtain feedback and take steps in the right direction.

Although the deconvolution approaches I tested all agreed on the principle contributions, these methods were fairly noisy when assigning contributions to low percentage proportions. Tissues that were inferred to be contributing less than 2% of overall cfDNA, for example adipose tissue, often exhibited high variance relative to their contribution across methods. Empirically, the exact numerical proportion assigned to a tissue seemed less reliable than the observed change from baseline, in this case our non-cancer controls. This emphasises the need for appropriate controls for comparative purposes in future deconvolution.

This also highlights a key factor missing from all current deconvolution methods, including my own: a reliable method for uncertainty quantification. The tissue contribution assignments made by these algorithms would be much more useful if each prediction was accompanied by a measure of confidence in the assignment. This is especially important to facilitate practical uptake of cfDNA deconvolution in clinical contexts, where the cost of a mis-assignment may be high. Alternatively, instead of assigning a single number, they could assign an interval set within which they are confident the true contribution lies. This can be integrated into neural network models using techniques such as conformal prediction [156]. The practical implementation of uncertainty quantification in supervised neural networks is a very active field of research, which may benefit my deep learning based deconvolution approach.

This highlights a more general advantage of re-framing cfDNA deconvolution and the multimodal classifier of cfDNA as supervised deep learning problems: they stand to benefit from the massive research effort in this space currently. My fairly simple method is easily extensible to new architectures or advances in explainable AI. Similarly, this supervised deconvolution method was trained on simulated data, the size of which could be easily scaled up. I noted improvements in model performance when doubling the size of the training set, and the other published model was trained on an even larger dataset and demonstrated even better performance. Future iterations of this deconvolution model could be scaled up dramatically to improve performance. The best approach would be to increase the diversity of the training data, simulating a wide range of biologically possible

scenarios, to ensure the trained model is robust to these scenarios. As we scale up size of training data, additional measures should be included to minimise risk of 'overfitting': learning patterns specific to the training data that do not generalise.

A consistent theme throughout this thesis is my attempt to devise approaches for overcoming the "curse of dimensionality" in clinical datasets: the general trend that these datasets typically have many more features than samples. This is exemplified by genome-wide analyses such as the cfDNA classification task faced here. Our objective is to separate liver cancer patients from non-cancer controls: we have 51 individuals in the cohort, and an entire genome worth of potentially informative features (around 28 million if we count each CpG separately). To counteract this inevitable challenge, we need to be maximally efficient with data use, and as far as possible train models using orthogonal data without using up precious patient data.

The fragment classifier I developed in chapter two was another example of reframing a small cohort classification problem as a task amenable to supervised deep learning. The key innovation there is to switch from classifying individual patients to classifying individual fragments. This of course does not replace the need for larger patient sample sizes to evaluate this model, but does provide a mechanism for more efficient data use. This approach worked well for bisulfite-sequenced cfDNA samples, but requires further optimisation and new training data to be useful for classification of TAPS-sequenced cfDNA. This was a key motivation for the construction of the tissue and tumour reference atlas for mC and hmC. This dataset will prove a valuable resource for training such models, and will be an immediate focus of the next steps in this project.

Alongside the numerous papers on cfDNA deconvolution and classification published in the last four years [72, 183, 101, 155], the fact that two other large reference atlases were published while this work was ongoing demonstrates the intense current interest in cfDNA classification and deconvolution. There is currently a heated competition to build the largest cfDNA dataset for cancer detection, and demonstrate the highest model sensitivity and specificity on this dataset. Large industry players are central to this competition, most notably the aptly-named company Grail. This race motivates a more thoughtful consideration of what is currently missing in cell-free DNA analysis.

Cell-free DNA has remained a fairly elusive biomolecule, and little is known about dynamics of tissue release, stability in the blood or degradation [152]. There remain large gaps in our understanding of the biology of this macromolecule, which could greatly aid such diagnostic models. For example, if we could obtain the tissue-specific rates of cfDNA shedding for each tissue in a healthy adult, we could take a more principled approach to cfDNA deconvolution, incorporating some of these biological priors. Additional benchmark datasets would also be very useful for cfDNA. These could be formed from experimental mixing of reads *in vitro* at a range of physiological concentrations, and could be used to evaluate cfDNA deconvolution and ctDNA classifier algorithms. Finally and more immediately, we could greatly improve such models by pooling resources and improving mechanisms for data sharing of cfDNA data, ideally at raw data stage to ensure uniform processing.

Larger training datasets unlock more complex models with more parameters and thus a greater capacity to learn. In many other domains, scaling up model training has resulted in vast improvements beyond expectation. The current example of large language models is a good example here. Training models to perform next-word prediction of text gives fairly unremarkable results on small scale data. However, by scaling model training up to a large proportion of the entire internet, and simultaneously increasing the model parameter size, resulting trained models are better than many experts predicted at a wide variety of tasks [109]. This surprisingly effective impact of simply scaling neural networks is known as "the bitter lesson" [184].

On the other hand, we should be cautious not to assume that scaling machine learning methods will solve many of these biological and clinical problems. Machine learning models are not a panacea here: the fundamental limitation for developing sensitive and specific cancer diagnostic tests is still data availability. High quality, publicly accessible data, with large sample sizes and free from batch processing effects remains the major bottleneck in advancing the field. When this is present, often simple, principled statistical approaches such as well-powered regressions have historically produced the most useful clinical results in cancer research [183, 72].

Relatedly, the goal of this research is ultimately to advance cancer research, and accurate diagnostic model performance is only a part of the puzzle here. It simply shows that there is some informative signal in the data which is successfully transferred by the modelling approach. There are many steps between obtaining a good performing model and getting a diagnostic biomarker adopted in the clinic. Practical details around implementation such as speed, cost and ease of use often limit uptake of novel computational diagnostic tools far more than performance scores. An accurate model is a necessary but not sufficient step towards non-invasive cancer diagnostics.

The availability of high-quality public data was a limitation for the cell-free DNA analyses described here. A key motivation for constructing our tissue reference atlas was to reduce such bottlenecks in future analyses. Our initial investigations represent a small fraction of the questions that can be asked of such a dataset, and by making this resource publicly available to the research community we can maximise its utility. The results I obtained from this preliminary analysis of the tissue atlas are promising, and support the notion that this will be an informative resource.

Firstly, judging by downstream results, the spatial segmentation method I developed was an effective compression of the 28 million CpGs across the genome into blocks of correlated mC and hmC. These can be considered functional units of methylation and hydroxymethylation across the genome, although further biological investigation is required to validate this claim. For example, we found that mC and hmC tissue specific blocks correlate well with histone markers denoting open chromatin, but the more general spatial relationship between these blocks and chromatin structure merits further study.

The tissue-specific DMBs identified here are both visually distinct, validated by downstream analysis, so are likely truly tissue informative regions. They are therefore good candidates for tissue-specific markers for subsequent analysis. The existence of these tissue-specific methylation markers, in particular hypo DMBs was not surprising, and has been confirmed multiple times by us and others [85, 144, 91]. Perhaps more surprising was the identification of tissue-specific hyper DMBs for hmC in both solid tissues and blood. These regions showed very little overlap with methylation DMBs, and were identified despite the presence

of brain in the reference comparison, which has a highly distinct hmC profile. These findings further establish hydroxymethylation as a distinct, tissue-specific epigenetic mark.

Tumour-specific differentially methylated blocks were most clearly distinct for hypo mC, although hyper DMBs for mC were visible for ovary, colon, liver and prostate tumour. Intriguingly, for a subset of tumours we also observe tumour-specific hyper DMBs for hydroxymethylation, whose role in cancer aetiology is largely unknown. Hydroxymethylation data was not available at the time of developing the cfDNA deconvolution or classification models. It was not until later that CAPS sequencing was validated and used for 5hmC sequencing, and this reference atlas is its first application. However, subsequent analyses will incorporate cfDNA sequenced using TAPSbeta and CAPS separately, to distinguish methylome from hydroxymethylome profiles, and the trained models will incorporate the hydroxymethylation tissue and tumour DMBs identified here.

An important caveat here is again around sample size and generalisability to new patients. Tumour samples are known to be highly heterogeneous, shaped by constant evolution and clonal expansion to be highly individualised. We must therefore exercise caution when grouping tumour types together, simply because they originate from the same tissue. The tumour specific signatures identified here may not generalise to other cancer samples from that tissue, and further validation of these DMBs on independent tumour data is needed. We attempted to maximise the number of tumour samples collected for each tissue in our reference, but again limitations of cost and data availability applied, and the compromise between depth and breadth in the atlas.

In the final chapter of this thesis, I begin to investigate the functional consequences of the tissue methylation and hydroxymethylation distributions we observe in this atlas. Gene expression is arguably the central determinant of cellular phenotype, so I decided this was a suitable prediction target. The correlations we observe between different functional genomic elements and gene expression largely agree with the literature, although a more fine-grained analysis is warranted for gene body mC and hmC, as this is too diverse an element to be accurately summarised with a single value. Another next step will be to investigate the gene

expression correlation across the tissue and tumour specific mC and hmC blocks I identified in Chapter 5, in both healthy and cancer tissue.

The predictive models I developed attempt to quantify the predictive power of mC and hmC for gene expression. The results we obtain on held out tissues are comparable to recent deep learning sequence-based models [104, 112]. However, because these sequence based models do not incorporate tissue-specific information, they may not be as useful for downstream tasks such as cancer-specific gene expression prediction.

7.0.1 Conclusions and key contributions

This thesis focuses on methods I developed to extract maximal information from genome-wide methylation data. In the first half I focus on their application to classifying cell-free DNA of cancer patients. Missing data is a key limitation of these results, and in the second half I focus on our efforts to rectify this by constructing a novel data resource: a tissue reference atlas of methylation and hydroxymethylation. I focus on what this novel data can tell us about the distribution and role of methylation and hydroxymethylation in healthy tissues, how this changes in cancer, and how this can be used to predict gene expression.

In Chapter 3, I focus on improving methods and data for tissue deconvolution of cell-free DNA. Previously limited to array-based deconvolution, I extend this approach to consider genome-wide tissue informative markers, by first collating the largest reference tissue atlas of genome-wide methylation signatures from public data. I then devise a simple algorithm for identifying tissue specific differentially methylated regions. I use this reference of genome-wide tissue informative markers to evaluate existing deconvolution methods on patient cell-free DNA, obtained from a pilot study of a novel epigenetic sequencing method. I demonstrate this approach gives meaningful biological results with diagnostic potential, and can be used to distinguish cfDNA of liver cancer patients from non-cancer controls. I then discuss the pitfalls of the current published tissue deconvolution methods, and introduce deconvolution method that uses supervised deep learning.

In Chapter 4, I devise methods of extracting more information from cell-free DNA through the use of multimodal classifiers. These classifiers better distinguish

cancer patients from non-healthy controls, thus promoting cfDNA utility as a non-invasive cancer biomarker. I demonstrate the information gain from incorporating multimodal information in an ensemble classifier, before developing a fragment-based approach to classify cfDNA that helps tackle the low sample size limitations of cancer cohort studies.

Missing data is a key limitation of experiments in Chapters 3 and 4, so in Chapter 5 I focus on our efforts to construct a novel data resource: a tissue reference atlas of methylation and hydroxymethylation. I describe the construction of this dataset, including the development of a novel genome spatial segmentation method to identify functional units of methylation and hydroxymethylation. I demonstrate the rich information that can be gleaned from an initial analyses, including identification of highly tissue-specific hydroxymethylation markers, and tumour-informative markers for mC and hmC. This dataset represents a valuable public resource to allow a more fine-grained analysis of mC and hmC, their interplay, and their influence on downstream cell function in healthy and cancer tissue.

In Chapter 6, I begin to characterise these downstream effects of mC and hmC by exploring their relationship with gene expression. I build a gene expression prediction model using these epigenetic markers, which demonstrates good performance in held-out tissues and independent cancer gene expression data. This model represents the first step in the process of examining the functional role of these two distinct epigenetic markers on gene expression, and how they can influence cellular phenotypes.

7.0.2 Future work

The cell-free DNA classifier models presented in Chapter 3 and 4 both suffer from scarcity of training data. This was a key motivation to construct our own tissue atlas containing matched tumour and healthy samples; to serve as training data for predictive models. An immediate next step is therefore to apply the tissue-specific markers identified in our atlas as tissue and tumour informative markers for model training. Given this rich training data, a particularly promising direction

is to improve our multimodal fragment classifier, which if tuned could sensitively detect small amounts of circulating tumour DNA in liquid biopsy.

Although the tumour-informative markers identified in this atlas appear robust and are validated with downstream correlation with gene expression and histone markers, further validation of these markers with public data is warranted. The tumour specific signatures identified here may not generalise to other cancer subtypes from that tissue, and further validation of these DMBs on independent tumour data is needed. At the time of writing, CAPS sequencing of cfDNA samples has just finished in our lab, which provides an additional layer of information to liquid biopsy: genome-wide hydroxymethylation signatures of cell-free DNA. Subsequent analyses will incorporate cfDNA sequenced using TAPSbeta and CAPS separately, to distinguish methylome from hydroxymethylome profiles, and the trained models will incorporate the hydroxymethylation tissue and tumour DMBs identified here.

The machine learning models presented in this thesis represent a preliminary investigation and a proof-of-concept; little time and resources was dedicated to optimising model architecture or scaling up model training, so it is possible that further improvements could be made here. An immediate next steps for models such as the cfDNA classifier, the deep autoencoder and the gene expression prediction model is therefore to further explore model hyperparameter tuning, and a more thorough examination of the data encoding and model choice, to ensure information transfer is maximised. Secondly, I shall investigate scaling up the models to test if we see performance gains as we increase number of parameters, whilst avoiding overfitting through regularisation techniques. Vitally, we must also include further validation with public data, such as independent gene expression data for the gene expression prediction model.

An independently validated gene expression prediction model could provide valuable biological insights into the role and importance of specific epigenetic features in this linkage, which can be probed using ablation experiments as is done for recent sequence-based gene prediction models [104]. This would be particularly interesting in tumour samples in our reference atlas, where we can begin to identify tumour-specific methylation and hydroxymethylation changes associated with cancer development. Finally, future iterations of this model could also be

adapted to predict gene expression programs given deconvoluted methylation and hydroxymethylation signatures observed in cell-free DNA, and thus be informative in a diagnostic model.

More broadly, there are many more avenues for exploration with regards to the interplay of methylation and hydroxymethylation in regulating gene expression, and how this changes in cancer. This tissue dataset, alongside the newly defined blocks and tissue informative markers for mC and hmC, will be a valuable resource for subsequent investigations into epigenetic changes in cancer. By combining this data with the approaches developed here for maximally efficient use of cfDNA and patient liquid biopsy data, new improved models can be built to help non-invasively diagnose cancer using these fascinating epigenetic markers.

Appendix A

Appendix

A.0.1 Cell-free DNA patient cohort

PDAC patients were consented for this study via the Oxford Radcliffe Biobank (09/H0606/5+5, project: 19/A177). Blood samples from patients with clinically diagnosed Pancreatitis, HCC and Cirrhosis patients were recruited for this trial at the John Radcliffe Hospital. The study was approved by Oxfordshire Research Ethics Committee A (REC-A) (10/H0604/51) and is registered on the U.K. National Institute for Health Research (NIHR) portfolio as study number 10776. Notably, the non-cancer controls were collected from an endoscopy clinic and were enriched with gastrointestinal inflammatory conditions such as Crohn's disease and colitis. Non-cancer control samples were collected from individuals without cancer diagnosis at the time of sample collection or previous history of cancer.

For all these patients, blood was collected into EDTA-coated Vacutainers, and plasma was separated from collected blood samples within 4 hours from collection. Cell-free DNA was extracted from plasma using the QIAamp Circulating Nucleic Acid Kit (Qiagen). The TAPS protocol was then used to label methylated cytosine, as described in [93], and TAPS libraries were constructed and paired-end 150 bp–sequenced on a NovaSeq 6000 sequencer (Illumina).

The age, sex and comorbidities patients in this cfDNA trial are summarised in Tables A.1 and A.2. The clinical metadata, stage and common blood biochemistry markers for each cancer patient in this cohort is listed in Table A.3.

Table A.1: Demographics and comorbidities in non-cancer patients

Sample	**Sex**	**Age**	**Batch**	**Comorbidities**
Ctrl_1	Female	53	Batch_1	CIBH, Ulcerative Colitis
Ctrl_2	Male	71	Batch_2	CIBH, Hypertension
Ctrl_3	Female	66	Batch_1	CIBH, Asthma, Bronchiectasis
Ctrl_4	Female	60	Batch_1	Hypertension
Ctrl_5	Male	58	Batch_3	Rectal bleeding
Ctrl_6	Female	46	Batch_3	NA
Ctrl_7	Male	79	Batch_4	Rectal bleeding, CIBH, Encephalitis
Ctrl_8	Female	65	Batch_2	Chronic diarrhoea, Diabetes
Ctrl_9	Male	69	Batch_3	Surveillance polyps, Epilepsy
Ctrl_10	Female	61	Batch_4	Hypertension
Ctrl_11	Male	48	Batch_1	IBDu, Ulcerative colitis
Ctrl_12	Male	69	Batch_2	Ulcerative Colitis
Ctrl_13	Female	45	Batch_2	Ulcerative Colitis, Distal proctitis
Ctrl_14	Female	45	Batch_4	TI focal active inflammation
Ctrl_15	Male	64	Batch_3	Crohn's disease, Possibly UC
Ctrl_16	Male	72	Batch_1	NA
Ctrl_17	Female	70	Batch_4	IBS, Sicca syndorme, Hypothyroidism
Ctrl_18	Female	68	Batch_1	PR bleeding
Ctrl_19	Female	65	Batch_3	Diabetes, Underactive thyroid, Hypertension
Ctrl_20	Male	78	Batch_3	AF, Hypertension
Ctrl_21	Male	50	Batch_2	Hypertension
Ctrl_22	Male	69	Batch_2	Duodenal polyp, Cardiomyopathy
Ctrl_23	Male	68	Batch_1	Duodenal polyp
Ctrl_24	Female	66	Batch_2	NA
Ctrl_25	Male	62	Batch_4	COPD, Hypertension
Ctrl_26	Male	81	Batch_1	IDA, Osteoarthritis, Hypopituituism
Ctrl_27	Female	48	Batch_4	CIBH, IDA
Ctrl_28	Female	75	Batch_4	Osteoarthritis, Raynauds, Diverticular disease
Ctrl_29	Male	72	Batch_4	Previous CABG, Aortic valve disease
Ctrl_30	Female	75	Batch_3	IDA, Hypertension, Angina
Cirrhosis_1	Male	40	Cirrhosis	NA
Cirrhosis_2	Female	61	Cirrhosis	NA
Cirrhosis_3	Male	74	Cirrhosis	NA
Cirrhosis_4	Male	45	Cirrhosis	NA
Pancreatitis_1	Female	81	Pancreatitis	Chronic pancreatitis with cysts, Diabetes
Pancreatitis_2	Male	53	Pancreatitis	Chronic calcific pancreatitis, Diabetes
Pancreatitis_3	Male	86	Pancreatitis	Diverticular disease, interstitial lung disease
Pancreatitis_4	Male	49	Pancreatitis	Anxiety
Pancreatitis_5	Female	35	Pancreatitis	NA
Pancreatitis_6	Female	35	Pancreatitis	Osteopenia
Pancreatitis_7	Male	37	Pancreatitis	Diverticulosis, pyloric stenosis

Table A.2: Demographics and comorbidities in cancer patients

Sample	Sex	Age	Batch	Comorbidities
HCC_1	Male	75	HCC	Unknown
HCC_2	Male	62	HCC	Unknown
HCC_3	Male	64	HCC	Unknown
HCC_4	Male	57	HCC	Unknown
HCC_5	Male	78	HCC	Unknown
HCC_6	Male	71	HCC	Unknown
HCC_7	Male	78	HCC	Unknown
HCC_8	Male	76	HCC	Unknown
HCC_9	Male	73	HCC	Unknown
HCC_10	Male	69	HCC	Unknown
HCC_11	Female	72	HCC	Unknown
HCC_12	Male	69	HCC	Unknown
HCC_13	Male	67	HCC	Unknown
HCC_14	Male	71	HCC	Unknown
HCC_15	Male	73	HCC	Unknown
HCC_16	Male	74	HCC	Unknown
HCC_17	Male	89	HCC	Unknown
HCC_18	Male	85	HCC	Unknown
HCC_19	Male	62	HCC	Unknown
HCC_20	Male	62	HCC	Unknown
HCC_21	Male	67	HCC	Unknown
PDAC_1	Female	78	PDAC	Hupertension, COPD
PDAC_2	Male	60	PDAC	Diabetes mellitus
PDAC_3	Male	78	PDAC	Diabetes mellitus, TIA, AF
PDAC_4	Male	75	PDAC	Acute kidney injury (Stage 1), chronic pancreatitis
PDAC_5	Male	78	PDAC	Left partial knee replacement, chronic pancreatitis
PDAC_6	Male	73	PDAC	Benign prostatic hypertrophy, new diagnosis of diabetes
PDAC_7	Female	79	PDAC	NA
PDAC_8	Female	59	PDAC	NA
PDAC_9	Female	80	PDAC	Osteoarthritis
PDAC_10	Male	62	PDAC	Porphyria, hypertension, IBS, sciatica
PDAC_11	Male	76	PDAC	Atrial fibrilation, Diabetes Mellitus (Type 2), Hypertension.
PDAC_12	Female	76	PDAC	Diabetes mellitus.
PDAC_13	Female	70	PDAC	NA
PDAC_14	Male	71	PDAC	NA
PDAC_15	Male	66	PDAC	Diabetes mellitus
PDAC_16	Female	74	PDAC	Previous breast and ovarian ca. brca 2 gene
PDAC_17	Male	66	PDAC	Hypertension
PDAC_18	Male	67	PDAC	NA
PDAC_19	Female	75	PDAC	Polymyalgia rheumatica, IBS, hysterectomy
PDAC_20	Female	66	PDAC	NA
PDAC_21	Male	71	PDAC	COPD, cardiac stents
PDAC_22	Male	57	PDAC	NA
PDAC_23	Male	69	PDAC	MI

Table A.3: Clinical metadata for HCC and PDAC patients

Sample	**Stage**	**Tumour size (mm)**	**Etiology**	**Cirrhosis**	**Child's Pugh**	**CA19-9 (U/mL)**	**AFP (KU/L)**
HCC_1	IIIA	130	No CLD	No	5	NA	1
HCC_2	I	27	ALD	Yes	6	NA	6
HCC_3	II	25	ALD	Yes	5	NA	4
HCC_4	I	23	NAFLD	Yes	5	NA	4
HCC_5	IVB	160	None	No	5	NA	50,000
HCC_6	III	70	NAFLD	Yes	11	NA	2
HCC_7	IIIA	87	NAFLD/ALD	Yes	5	NA	8
HCC_8	II	75	No CLD	No	5	NA	5
HCC_9	II	25	NAFLD	Yes	7	NA	5
HCC_10	I	50	NAFLD	No	5	NA	3
HCC_11	II	30	PBC	No	5	NA	3
HCC_12	II	17	MS	Yes	5	NA	4
HCC_13	I	16	MS	Yes	5	NA	5
HCC_14	I	12	Mixed ALD/MS	Yes	5	NA	3
HCC_15	II	28	NAFLD	Yes	5	NA	30
HCC_16	I	54	NAFLD	No	5	NA	2
HCC_17	I	24	NAFLD	No	5	NA	1087
HCC_18	IIIA	118	No clear CLD	No	5	NA	9
HCC_19	IIIB	30	MS	Yes	5	NA	16
HCC_20	I	9	ALD	Yes	5	NA	118
HCC_21	IIIA	33	NAFLD	Yes	5	NA	11
PDAC_1	III	28	NA	NA	NA	NA	NA
PDAC_2	IIB	40	NA	NA	NA	41	NA
PDAC_3	III	45	NA	NA	NA	6166	NA
PDAC_4	III	42	NA	NA	NA	NA	NA
PDAC_5	III	30	NA	NA	NA	240	NA
PDAC_6	IV	40	NA	NA	NA	NA	NA
PDAC_7	IIB	36	NA	NA	NA	432	NA
PDAC_8	III	40	NA	NA	NA	NA	NA
PDAC_9	IIB	60	NA	NA	NA	130	NA
PDAC_10	IB	27	NA	NA	NA	338	NA
PDAC_11	III	30	NA	NA	NA	2,006	NA
PDAC_12	III	38	NA	NA	NA	510	NA
PDAC_13	IIB	30	NA	NA	NA	154	NA
PDAC_14	III	45	NA	NA	NA	2250	NA
PDAC_15	IIB	60	NA	NA	NA	1534	NA
PDAC_16	IA	18	NA	NA	NA	27	NA
PDAC_17	III	30	NA	NA	NA	140	NA
PDAC_18	IIB	24	NA	NA	NA	51	NA
PDAC_19	IA	28	NA	NA	NA	88	NA
PDAC_20	IIB	12	NA	NA	NA	8	NA
PDAC_21	III	40	NA	NA	NA	2000	NA
PDAC_22	IIB	35	NA	NA	NA	NA	NA
PDAC_23	III	35	NA	NA	NA	900	NA

A.0.2 TAPS tissue atlas QC and clustering

Tissue samples in our atlas were originally collected through surgeries or autopsies. Immune cell populations were isolated from blood taken from healthy donors by serial purifications by magnetic beads. Flow cytometry confirmed a high purity of over 80% for the majority of the immune cell populations. Megakaryocytes and erythroid precursors were in vitro differentiated from CD34+ cells from mobilised healthy donors, and we collected pure cell populations by FACS. We obtained DNA from these tissue and blood cell samples for the atlas, so do not know the original tissue size. This was obtained from a variety of sources: the majority were purchased online through various vendors, except for blood and immune cell types which were sourced from collaborators. Table A.4 shows the source of each sample in our atlas, the age and sex of the donor and the tumour contribution for the tumour tissues.

We performed TAPS and CAPS+ on the 116 samples to obtain whole-genome and base-resolution methylome and hydroxymethylome of these samples. We achieved mean depths of 30.6±4.1x and 29.7±4.6x at CpG sites for TAPS and CAPS+, respectively. The conversion rates analysed on the spike-ins demonstrated high conversion rates of target modifications and low false positive rates on the other cytosine modifications and unmodified cytosines, suggesting TAPS and CAPS+ datasets captured true methylome and hydroxymethylome. To ensure accurate measurements of 5mC and 5hmC, we filtered out CpG sites that are covered less than 3x in any samples. The resulting atlas covers 23,249,723 and 23,711,956 CpG sites for 5mC and 5hmC, respectively.

All donors provided informed, written consent. Genomic DNA samples purchased from Origene and AMSBIO were collected under ethics approvals by Institutional Review Board (IRB). For erythroid precursors and megakaryocytes, samples were banked and stored under the approved InForMED Study protocol (IRAS 199833; REC 16/Lo/1376). For the rest of the blood cells, the ethical approval was granted by London - Bloomsbury Research Ethics Committee (REC ref: 21/PR/1155).

Table A.4: Clinical information for the samples included in the atlas. % Lesional and % Tumour are based on pathological analysis.

Sample type	Source	Sample/donor ID	Gender	Age	Race	% Lesional	% Tumour
Brain	AMSBIO	C505016	Male	77	White	NA	NA
Brain	AMSBIO	C707004	Female	78	White	NA	NA
Brain-Tumour	AMSBIO	C707008	Female	60	Asian	NA	NA
Brain-Tumour	AMSBIO	C707009	Female	63	Asian	NA	NA
Breast	Origene	CD563137	Female	61	White	0	0
Breast	Origene	CD563304	Female	64	White	0	0
Breast	Origene	CD564068	Female	63	NA	0	0
Breast	Origene	CD564368	Female	76	NA	0	0
Breast-Tumour	Origene	CD563503	Female	70	White	0	90
Breast-Tumour	Origene	CD564076	Female	63	NA	0	90
Breast-Tumour	Origene	CD564575	Female	72	NA	0	95
Breast-Tumour	Origene	CD565056	Female	77	White	0	90
Colon	Origene	CD563419	Male	61	White	0	0
Colon	Origene	CD563663	Male	75	White	0	0
Colon	Origene	CD564159	Female	76	White	0	0
Colon	Origene	CD565189	Female	77	NA	0	0
Colon-Tumour	Origene	CD563649	Female	66	Black	0	70
Colon-Tumour	Origene	CD564145	Male	72	White	0	90
Colon-Tumour	Origene	CD564971	Male	66	NA	0	100
Colon-Tumour	Origene	CD565203	Female	74	NA	0	75
Esophagus	Origene	CD563678	Male	60	White	0	0
Esophagus	Origene	CD564986	Male	63	White	0	0
Esophagus	Origene	CD565136	Male	76	NA	0	0
Esophagus	Origene	CD565252	Male	73	NA	0	0
Heart	AMSBIO	C707006	Male	71	White	NA	NA
Kidney	Origene	CD563776	Female	63	NA	0	0
Kidney	Origene	CD563778	Male	77	NA	0	0
Kidney	Origene	CD563984	Male	71	NA	0	0
Kidney	Origene	CD564241	Female	76	White	0	0
Kidney-Tumour	Origene	CD563857	Male	61	White	0	97
Kidney-Tumour	Origene	CD564876	Female	76	NA	0	98
Kidney-Tumour	Origene	CD564953	Male	71	NA	0	98
Kidney-Tumour	Origene	CD565153	Female	64	NA	0	90
Liver	Origene	CD563257	Female	69	White	0	0
Liver	Origene	CD564082	Male	73	White	0	0
Liver	Origene	CD564797	Male	72	NA	0	0
Liver-Cirrhosis	Origene	CD563106	Male	71	White	100	0
Liver-Cirrhosis	Origene	CD563278	Male	43	White	100	0
Liver-Cirrhosis	Origene	CD564801	Female	58	NA	100	0
Liver-Cirrhosis	Origene	CD564847	Male	56	NA	100	0
Liver-Tumour	Origene	CD563176	Male	63	NA	0	90
Liver-Tumour	Origene	CD564146	Male	68	White	0	98
Liver-Tumour	Origene	CD564208	Male	73	NA	0	95
Liver-Tumour	Origene	CD564934	Male	71	NA	0	90
Lung	Origene	CD563854	Female	65	White	0	0
Lung	Origene	CD563930	Female	72	White	0	0
Lung	Origene	CD564511	Male	72	NA	0	0
Lung	Origene	CD564901	Male	67	NA	0	0
Lung-Tumour	Origene	CD564285	Male	63	Black	0	80
Lung-Tumour	Origene	CD564572	Male	70	NA	0	95
Lung-Tumour	Origene	CD565099	Female	79	Black	0	80
Lung-Tumour	Origene	CD565285	Female	61	NA	0	90
Ovary	Origene	CD563544	Female	54	White	0	0
Ovary	Origene	CD564191	Female	63	Black	0	0
Ovary	Origene	CD564295	Female	74	White	0	0
Ovary-Tumour	Origene	CD563808	Female	80	Black	0	85
Ovary-Tumour	Origene	CD564888	Female	60	NA	0	85
Ovary-Tumour	Origene	CD565018	Female	62	White	0	80
Ovary-Tumour	Origene	CD565088	Female	75	White	0	80
Pancreas	Origene	CD564011	Male	71	NA	0	0
Pancreas	Origene	CD564404	Male	57	NA	0	0
Pancreas	Origene	CD564844	Male	66	NA	0	0
Pancreas	Origene	CD565341	Male	61	NA	0	0
Pancreas-Pancreatitis	Origene	CD563553	Male	75	White	100	0
Pancreas-Tumour	Origene	CD564127	Male	75	White	0	30
Pancreas-Tumour	Origene	CD564422	Male	57	NA	0	85
Prostate	Origene	CD563267	Male	61	NA	0	0
Prostate	Origene	CD563610	Male	63	White	0	0
Prostate	Origene	CD563685	Male	71	Black	0	0
Prostate	Origene	CD564242	Male	67	White	0	0
Prostate-Tumour	Origene	CD563761	Male	66	White	0	90
Prostate-Tumour	Origene	CD563882	Male	60	NA	0	90
Prostate-Tumour	Origene	CD564414	Male	62	NA	0	95

Continued on next page

Table A.4 – continued from previous page

Sample type	Source	Sample/donor ID	Gender	Age	Race	% Lesional	% Tumour
Prostate-Tumour	Origene	CD564980	Male	65	White	0	90
Spleen	Origene	CD563880	Female	66	White	0	0
Spleen	Origene	CD565017	Male	65	Black	0	0
Stomach	Origene	CD563162	Male	74	NA	0	0
Stomach	Origene	CD563430	Male	67	White	0	0
Stomach	Origene	CD564596	Male	77	NA	0	0
Stomach	Origene	CD565042	Male	71	White	0	0
Stomach-Tumour	Origene	CD564588	Male	77	NA	0	55
Stomach-Tumour	Origene	CD564659	Female	67	NA	0	85
B-cells	Dong Group	Donor 1	Female	56	Asian	NA	NA
B-cells	Dong Group	Donor 2	Male	62	White	NA	NA
B-cells	Dong Group	Donor 3	Male	55	White	NA	NA
B-cells	Dong Group	Donor 4	Female	55	Asian	NA	NA
CD4-T-cells	Dong Group	Donor 1	Female	56	Asian	NA	NA
CD4-T-cells	Dong Group	Donor 2	Male	62	White	NA	NA
CD4-T-cells	Dong Group	Donor 3	Male	55	White	NA	NA
CD4-T-cells	Dong Group	Donor 4	Female	55	Asian	NA	NA
CD8-T-cells	Dong Group	Donor 1	Female	56	Asian	NA	NA
CD8-T-cells	Dong Group	Donor 2	Male	62	White	NA	NA
CD8-T-cells	Dong Group	Donor 3	Male	55	White	NA	NA
CD8-T-cells	Dong Group	Donor 4	Female	55	Asian	NA	NA
Eosinophils	Dong Group	Donor 1	Female	56	Asian	NA	NA
Eosinophils	Dong Group	Donor 2	Male	62	White	NA	NA
Eosinophils	Dong Group	Donor 3	Male	55	White	NA	NA
Eosinophils	Dong Group	Donor 4	Female	55	Asian	NA	NA
Monocytes	Dong Group	Donor 1	Female	56	Asian	NA	NA
Monocytes	Dong Group	Donor 2	Male	62	White	NA	NA
Monocytes	Dong Group	Donor 3	Male	55	White	NA	NA
Monocytes	Dong Group	Donor 4	Female	55	Asian	NA	NA
Neutrophils	Dong Group	Donor 1	Female	56	Asian	NA	NA
Neutrophils	Dong Group	Donor 2	Male	62	White	NA	NA
Neutrophils	Dong Group	Donor 3	Male	55	White	NA	NA
Neutrophils	Dong Group	Donor 4	Female	55	Asian	NA	NA
NK-cells	Dong Group	Donor 1	Female	56	Asian	NA	NA
NK-cells	Dong Group	Donor 2	Male	62	White	NA	NA
NK-cells	Dong Group	Donor 3	Male	55	White	NA	NA
NK-cells	Dong Group	Donor 4	Female	55	Asian	NA	NA
Erythroid-precursors	Psaila Group	Donor 30	NA	NA	NA	NA	NA
Erythroid-precursors	Psaila Group	Donor 37	NA	NA	NA	NA	NA
Erythroid-precursors	Psaila Group	Donor 66	Male	NA	NA	NA	NA
Erythroid-precursors	Psaila Group	Donor 70	Male	NA	NA	NA	NA
Megakaryocytes	Psaila Group	Donor 30	NA	NA	NA	NA	NA
Megakaryocytes	Psaila Group	Donor 66	Male	NA	NA	NA	NA

The tumour types in our atlas were from single specific cancer subtypes, as listed in Table A.5.

Table A.5: Tumour types included in the atlas

Sample type	**Tumour type**
Brain-Tumour	Glioblastoma
Breast-Tumour	Adenocarcinoma of breast
Colon-Tumour	Adenocarcinoma of colon
Kidney-Tumour	Clear cell renal cell carcinoma
Liver-Tumour	Hepatocellular carcinoma
Lung-Tumour	Adenocarcinoma of lung
Ovary-Tumour	Ovarian serous adenocarcinoma
Pancreas-Tumour	Adenocarcinoma of pancreas
Prostate-Tumour	Adenocarcinoma of prostate
Stomach-Tumour	Adenocarcinoma of stomach

Before incorporating into our reference atlas, a number of quality control (QC) checks were necessary to judge suitability of each tissue sample for inclusion. The following criteria was used for each sample:

1. Sufficient depth of coverage
2. Even read coverage across the genome
3. Expected methylation ratio distribution
4. No batch effect, clustering near other samples from the same tissue

All tissue samples had a sufficiently high mean coverage depth above 20x, with a global mean of 30.76x across all TAPSbeta samples and 30.12x across all CAPS samples, as illustrated in Figure 5.2. However, this high mean coverage belies some outlier samples which have a high proportion of non-covered CpG sites. These outlier samples exhibited a very uneven coverage, where numerous CpG sites are not covered at all by sequencing, but the remaining CpG sites which are covered have high average depth. We can identify these samples by counting the number of CpG sites with zero coverage in each sample, or coverage less than some threshold, set to 4 in Figure A.3.

Subsequent CpG-level analysis relies on obtaining a tissue reference matrix of unique CpG sites across the genome, where *all CpG sites* are covered above some threshold depth in *all* samples. If a single tissue sample has zero coverage or less than the threshold coverage at a specific CpG, we have three options:

1. Discard that CpG locus across all samples, thus reducing breadth of genome covered
2. Discard that tissue sample altogether, thus reducing our tissue sample size
3. Impute the methylation or hydroxymethylation ratio for that tissue at that CpG locus.

We want to minimise occurence of 1 and 2 as much as possible, to keep the proportion of the genome covered as high as possible, and to retain the largest possible sample size respectively. I explored imputation, but it presents several obstacles: a) it is not reliable for tissue groups with just one or two replicates, b) it fails to acccount for genuine outliers and artificially pulls these samples towards the mean, and c) it requires many parameters to be set, such as degree of missingness. Ultimately imputation can obscure the original data and falsely reduce true biological variability, which is not desirable here.

A simpler solution is to identify and remove the problematic samples with uneven coverage. Further analysis of the copy number aberrations in these outlier samples confirmed they're highly unusual genomic coverage distribution: as shown in Figure A.2: one heart sample has normal evenly distributed coverage, whilst a second has highly uneven coverage. I therefore elected to remove these 4 outlier samples from all subsequent analysis.

The methylation ratio at each CpG is an imperfect representation of the underlying methylation state at that locus, represented by total and methylated read counts. An important parameter in this QC stage is the threshold read count at each CpG, below which we refuse to trust the calculated methylation ratio at that CpG and discard it. By setting this too low, we risk including methylation ratios at certain CpGs which are not a true reflection of the underlying methylation state, thus introducing bias into our reference atlas. Subsequent analysis relies

on all samples being covered above the selected threshold at all CpG sites by all samples. By setting this depth threshold too high, we lose excessive CpGs. We set the coverage threshold at 3 for all subsequent analysis. This trade-off is illustrated in Figure A.1

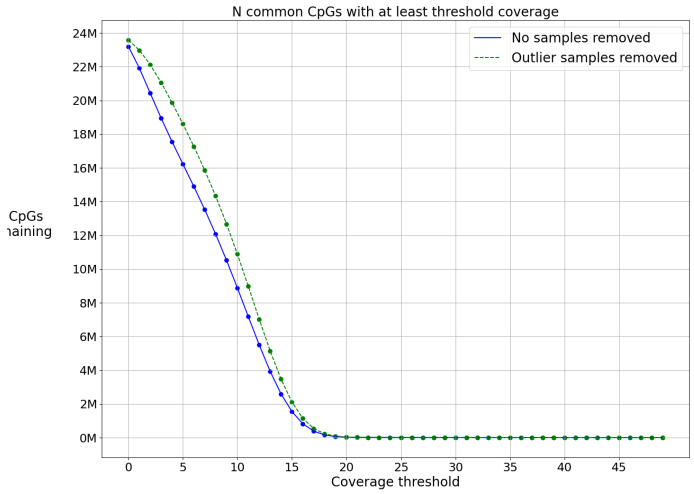


Figure A.1: Number of remaining CpGs after filtering all rows of the reference atlas by different coverage thresholds.

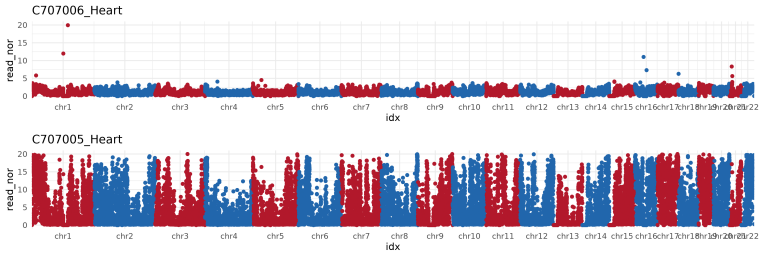


Figure A.2: Copy number variation plot for two heart samples generated by ichorCNA. The upper panel is representative of a 'normal' CNA plot, whereas the lower plot shows wide heterogeneity indicating sample quality issues.

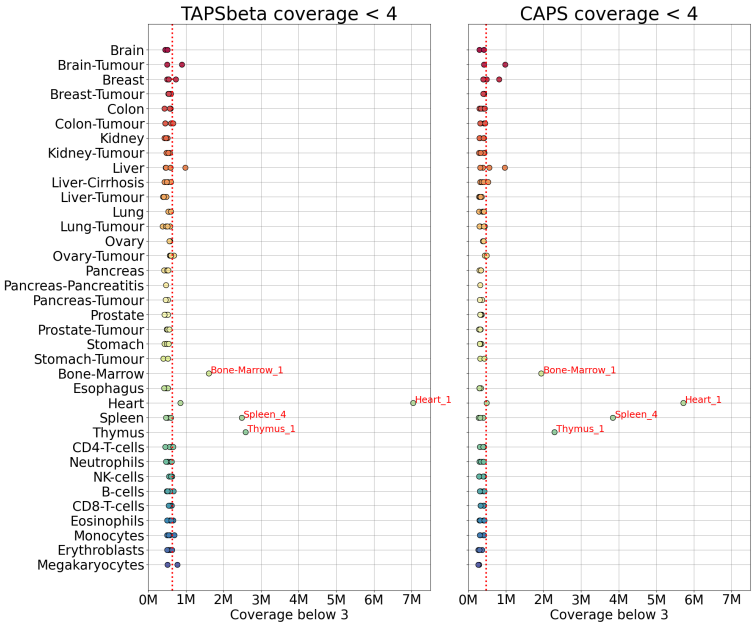


Figure A.3: Count of CpG sites in each sample with coverage less than 4. Outlier samples with many low coverage sites are labelled in red.

After removing outlier by anomalous coverage, I turned to overall methylation and hydroxymethylation signatures across tissues, and between tissue samples within the same tissue group. Broadly the question we're asking is: to what extent do we see tissue specific methylation and hydroxymethylation, and are there any outliers with an unusual methylation signature for their tissue group. I used three lines of enquiry:

1. Inter- and intra- tissue correlation in methylation and hydroxymethylation
2. Dimensionality reduction and visualisation of all tissue samples
3. Deconvolution of tissue samples using our previous reference

To help identify outliers, I performed NNLS tissue deconvolution on these new tissue samples using a recently published cell-type methylation atlas as a reference (Figure A.4).

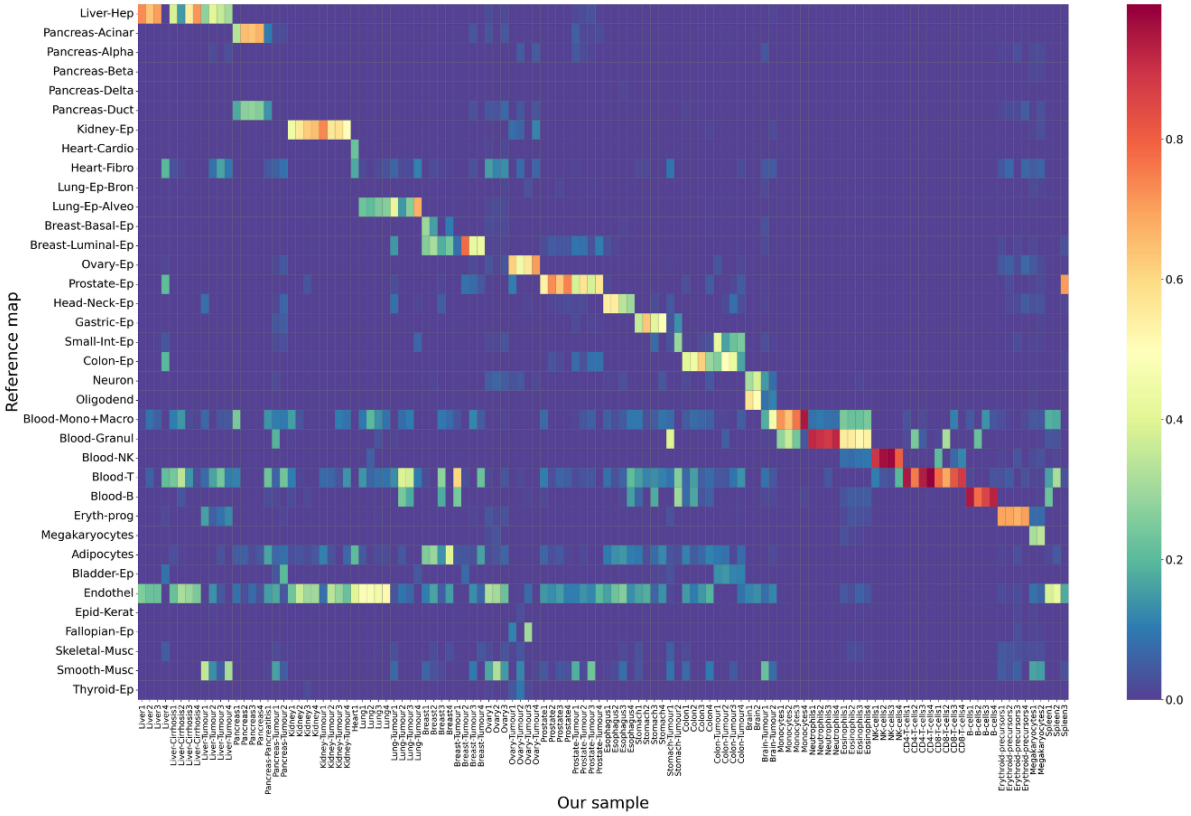


Figure A.4: Deconvolution of our reference atlas on a recently published methylation reference atlas. There is not a one-to-one correspondence between tissue types in the two references, but they have been matched as closely as possible.

Bibliography

[1] Stephen B. Baylin and Peter A. Jones. A decade of exploring the cancer epigenome — biological and translational implications. *Nature Reviews. Cancer*, 11(10):726–734, September 2011.

[2] Yuval Dor and Howard Cedar. Principles of DNA methylation and their implications for biology and medicine. *The Lancet*, 392(10149):777–786, September 2018.

[3] Daniel D De Carvalho, Jueng Soo You, and Peter A Jones. DNA Methylation and Cellular Reprogramming. *Trends in cell biology*, 20(10):609–617, October 2010.

[4] Oswald T. Avery, Colin M. MacLeod, and Maclyn McCarty. STUDIES ON THE CHEMICAL NATURE OF THE SUBSTANCE INDUCING TRANSFORMATION OF PNEUMOCOCCAL TYPES. *The Journal of Experimental Medicine*, 79(2):137–158, February 1944.

[5] R. Holliday and J. E. Pugh. DNA modification mechanisms and gene activity during development. *Science (New York, N.Y.)*, 187(4173):226–232, January 1975.

[6] A. Bird, M. Taggart, M. Frommer, O. J. Miller, and D. Macleod. A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. *Cell*, 40(1):91–99, January 1985.

[7] Yimeng Yin, Ekaterina Morgunova, Arttu Jolma, Eevi Kaasinen, Biswajyoti Sahu, Syed Khund-Sayeed, Pratyush K. Das, Teemu Kivioja, Kashyap Dave, Fan Zhong, Kazuhiro R. Nitta, Minna Taipale, Alexander Popov, Paul A.

Ginno, Silvia Domcke, Jian Yan, Dirk Schübeler, Charles Vinson, and Jussi Taipale. Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science*, 356(6337):eaaj2239, May 2017. Publisher: American Association for the Advancement of Science.

[8] J. Boyes and A. Bird. DNA methylation inhibits transcription indirectly via a methyl-CpG binding protein. *Cell*, 64(6):1123–1134, March 1991.

[9] Lisa D. Moore, Thuc Le, and Guoping Fan. DNA Methylation and Its Basic Function. *Neuropsychopharmacology*, 38(1):23–38, January 2013. Number: 1 Publisher: Nature Publishing Group.

[10] Robert S. Illingworth, Ulrike Gruenewald-Schneider, Shaun Webb, Alastair R. W. Kerr, Keith D. James, Daniel J. Turner, Colin Smith, David J. Harrison, Robert Andrews, and Adrian P. Bird. Orphan CpG Islands Identify Numerous Conserved Promoters in the Mammalian Genome. *PLoS Genetics*, 6(9):e1001134, September 2010.

[11] Paige A. Bommarito and Rebecca C. Fry. Chapter 2-1 - The Role of DNA Methylation in Gene Regulation. In Shaun D. McCullough and Dana C. Dolinoy, editors, *Toxicoepigenetics*, pages 127–151. Academic Press, January 2019.

[12] Peter A. Jones. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Reviews Genetics*, 13(7):484–492, July 2012. Number: 7 Publisher: Nature Publishing Group.

[13] Michael J. Ziller, Hongcang Gu, Fabian Müller, Julie Donaghey, Linus T.-Y. Tsai, Oliver Kohlbacher, Philip L. De Jager, Evan D. Rosen, David A. Bennett, Bradley E. Bernstein, Andreas Gnirke, and Alexander Meissner. Charting a dynamic DNA methylation landscape of the human genome. *Nature*, 500(7463):477–481, August 2013. Number: 7463 Publisher: Nature Publishing Group.

[14] Rafael A. Irizarry, Christine Ladd-Acosta, Bo Wen, Zhijin Wu, Carolina Montano, Patrick Onyango, Hengmi Cui, Kevin Gabo, Michael Rongione,

Maree Webster, Hong Ji, James Potash, Sarven Sabunciyan, and Andrew P. Feinberg. Genome-wide methylation analysis of human colon cancer reveals similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nature genetics*, 41(2):178–186, February 2009.

[15] Michael Weber, Jonathan J. Davies, David Wittig, Edward J. Oakeley, Michael Haase, Wan L. Lam, and Dirk Schübeler. Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nature Genetics*, 37(8):853–862, August 2005. Number: 8 Publisher: Nature Publishing Group.

[16] F.M. Smith, A.S. Garfield, and A. Ward. Regulation of growth and metabolism by imprinted genes. *Cytogenetic and Genome Research*, 113(1-4):279–291, March 2006.

[17] Wei Xie, Cathy L. Barr, Audrey Kim, Feng Yue, Ah Young Lee, James Eubanks, Emma L. Dempster, and Bing Ren. Base-resolution analyses of sequence and parent-of-origin dependent DNA methylation in the mouse genome. *Cell*, 148(4):816–831, February 2012.

[18] Martina Barchitta, Annalisa Quattrocchi, Andrea Maugeri, Manlio Vinciguerra, and Antonella Agodi. LINE-1 Hypomethylation in Blood and Tissue Samples as an Epigenetic Marker for Cancer Risk: A Systematic Review and Meta-Analysis. *PLOS ONE*, 9(10):e109478, October 2014. Publisher: Public Library of Science.

[19] S. Kochanek, D. Renz, and W. Doerfler. DNA methylation in the Alu sequences of diploid and haploid primary human cells. *The EMBO Journal*, 12(3):1141–1151, March 1993. Num Pages: 1151 Publisher: John Wiley & Sons, Ltd.

[20] Miho M. Suzuki and Adrian Bird. DNA methylation landscapes: provocative insights from epigenomics. *Nature Reviews Genetics*, 9(6):465–476, June 2008. Number: 6 Publisher: Nature Publishing Group.

[21] N. Shenker and J. M. Flanagan. Intragenic DNA methylation: implications of this epigenetic mechanism for cancer research. *British Journal of Cancer*, 106(2):248–253, January 2012. Number: 2 Publisher: Nature Publishing Group.

[22] Michael B. Stadler, Rabih Murr, Lukas Burger, Robert Ivanek, Florian Lienert, Anne Schöler, Erik van Nimwegen, Christiane Wirbelauer, Edward J. Oakeley, Dimos Gaidatzis, Vijay K. Tiwari, and Dirk Schübeler. DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature*, 480(7378):490–495, December 2011. Number: 7378 Publisher: Nature Publishing Group.

[23] Jason Ernst and Manolis Kellis. Chromatin-state discovery and genome annotation with ChromHMM. *Nature Protocols*, 12(12):2478–2492, December 2017. Number: 12 Publisher: Nature Publishing Group.

[24] Allegra Angeloni and Ozren Bogdanovic. Enhancer DNA methylation: implications for gene regulation. *Essays in Biochemistry*, 63(6):707–715, September 2019.

[25] Emily Hodges, Antoine Molaro, Camila O. Dos Santos, Pramod Thekkat, Qiang Song, Philip J. Uren, Jin Park, Jason Butler, Shahin Rafii, W. Richard McCombie, Andrew D. Smith, and Gregory J. Hannon. Directional DNA Methylation Changes and Complex Intermediate States Accompany Lineage Specificity in the Adult Hematopoietic Compartment. *Molecular cell*, 44(1):17–28, October 2011.

[26] Xi Chen, Jian Zhou, Ran Zhang, Aaron K. Wong, Christopher Y. Park, Chandra L. Theesfeld, and Olga G. Troyanskaya. Tissue-specific enhancer functional networks for associating distal regulatory regions to disease. *Cell Systems*, 12(4):353–362.e6, April 2021.

[27] Beatrice Borsari, Pablo Villegas-Mirón, Sílvia Pérez-Lluch, Isabel Turpin, Hafid Laayouni, Alba Segarra-Casas, Jaume Bertranpetit, Roderic Guigó, and Sandra Acosta. Enhancers with tissue-specific activity are enriched in

intronic regions. *Genome Research*, July 2021. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

[28] Douglas Hanahan and Robert A. Weinberg. Hallmarks of Cancer: The Next Generation. *Cell*, 144(5):646–674, March 2011. Publisher: Elsevier.

[29] Douglas Hanahan. Hallmarks of Cancer: New Dimensions. *Cancer Discovery*, 12(1):31–46, January 2022.

[30] Edward S. Diala and Robert M. Hoffman. Hypomethylation of hela cell DNA and the absence of 5-methylcytosine in SV40 and adenovirus (type 2) DNA: Analysis by HPLC. *Biochemical and Biophysical Research Communications*, 107(1):19–26, July 1982.

[31] Andrew P. Feinberg and Bert Vogelstein. Hypomethylation distinguishes genes of some human cancers from their normal counterparts. *Nature*, 301(5895):89–92, January 1983. Number: 5895 Publisher: Nature Publishing Group.

[32] Manel Esteller, Jose Maria Silva, Gemma Dominguez, Felix Bonilla, Xavier Matias-Guiu, Enrique Lerma, Elena Bussaglia, Jaime Prat, I. Clara Harkes, Elizabeth A. Repasky, Edward Gabrielson, Mieke Schutte, Stephen B. Baylin, and James G. Herman. Promoter Hypermethylation and BRCA1 Inactivation in Sporadic Breast and Ovarian Tumors. *JNCI: Journal of the National Cancer Institute*, 92(7):564–569, April 2000.

[33] J G Herman, F Latif, Y Weng, M I Lerman, B Zbar, S Liu, D Samid, D S Duan, J R Gnarra, and W M Linehan. 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):9700–9704, October 1994.

[34] M. Esteller, R. A. Risques, M. Toyota, G. Capella, V. Moreno, M. A. Peinado, S. B. Baylin, and J. G. Herman. Promoter hypermethylation of

the DNA repair gene O(6)-methylguanine-DNA methyltransferase is associated with the presence of G:C to A:T transition mutations in p53 in human colorectal tumorigenesis. *Cancer Research*, 61(12):4689–4692, June 2001.

[35] Srinivasan Yegnasubramanian, Michael C. Haffner, Yonggang Zhang, Bora Gurel, Toby C. Cornish, Zhijin Wu, Rafael Irizarry, James Morgan, Jessica Hicks, Theodore L. DeWeese, William B. Isaacs, G. Steven Bova, Angelo M. De Marzo, and William G. Nelson. DNA hypomethylation arises later in prostate cancer progression than CpG island hypermethylation and contributes to metastatic tumor heterogeneity. *Cancer research*, 68(21):8954–8967, November 2008.

[36] Szilvia I Ecsedi, Hector Hernandez-Vargas, Sheila C Lima, Zdenko Herceg, Roza Adany, and Margit Balazs. Transposable hypomethylation is associated with metastatic capacity of primary melanomas. *International Journal of Clinical and Experimental Pathology*, 6(12):2943–2948, November 2013.

[37] Signe Søes, Iben Lyster Daugaard, Brita Singers Sørensen, Andreas Carus, Manuel Mattheisen, Jan Alsner, Jens Overgaard, Henrik Hager, Lise Lotte Hansen, and Lasse Sommer Kristensen. Hypomethylation and increased expression of the putative oncogene ELMO3 are associated with lung cancer development and metastases formation. *Oncoscience*, 1(5):367–374, May 2014.

[38] Jinchang Lu, Guohui Song, Qinglian Tang, Changye Zou, Feng Han, Zhiqiang Zhao, Bicheng Yong, Junqiang Yin, Huaiyuan Xu, Xianbiao Xie, Tiebang Kang, YingLee Lam, Huiling Yang, Jingnan Shen, and Jin Wang. IRX1 hypomethylation promotes osteosarcoma metastasis via induction of CXCL14/NF-KB signaling. *The Journal of Clinical Investigation*, 125(5):1839–1856, May 2015.

[39] Ranjani Lakshminarasimhan and Gangning Liang. The Role of DNA Methylation in Cancer. *Advances in experimental medicine and biology*, 945:151–172, 2016.

[40] Skirmantas Kriaucionis and Nathaniel Heintz. The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. *Science (New York, N.Y.)*, 324(5929):929–930, May 2009.

[41] Maxim Ivanov, Mart Kals, Marina Kacevska, Isabel Barragan, Kie Kasuga, Anders Rane, Andres Metspalu, Lili Milani, and Magnus Ingelman-Sundberg. Ontogeny, distribution and potential roles of 5-hydroxymethylcytosine in human liver function. *Genome Biology*, 14(8):R83, 2013.

[42] Marian Mellen, Pinar Ayata, Scott Dewell, Skirmantas Kriaucionis, and Nathaniel Heintz. MeCP2 binds to 5hmc enriched within active genes and accessible chromatin in the nervous system. *Cell*, 151(7):1417–1430, December 2012.

[43] Lin-Yong Zhao, Jinghui Song, Yibin Liu, Chun-Xiao Song, and Chengqi Yi. Mapping the epigenetic modifications of DNA and RNA. *Protein & Cell*, 11(11):792–808, November 2020.

[44] Ageliki Tsagaratou, Tarmo Äijö, Chan-Wang J. Lio, Xiaojing Yue, Yun Huang, Steven E. Jacobsen, Harri Lähdesmäki, and Anjana Rao. Dissecting the dynamic changes of 5-hydroxymethylcytosine in T-cell development and differentiation. *Proceedings of the National Academy of Sciences of the United States of America*, 111(32):E3306–E3315, August 2014.

[45] Lucile Marion-Poll, Jean-Pierre Roussarie, Lieng Taing, Cloelia Dard-Dascot, Nicolas Servant, Yan Jaszczyszyn, Emmanuelle Jordi, Eskeatnaf Mulugeta, Denis Hervé, Déborah Bourc'his, Paul Greengard, Claude Thermes, and Jean-Antoine Girault. DNA methylation and hydroxymethylation characterize the identity of D1 and D2 striatal projection neurons. *Communications Biology*, 5:1321, December 2022.

[46] Yibin Liu, Zhiyuan Hu, Jingfei Cheng, Paulina Siejka-Zielińska, Jinfeng Chen, Masato Inoue, Ahmed Ashour Ahmed, and Chun-Xiao Song. Subtraction-free and bisulfite-free specific sequencing of 5-methylcytosine

and its oxidized derivatives at base resolution. *Nature Communications*, 12(1):618, January 2021. Number: 1 Publisher: Nature Publishing Group.

[47] Rajneesh Richa and Rajeshwar P. Sinha. Hydroxymethylation of DNA: an epigenetic marker. *EXCLI Journal*, 13:592–610, May 2014.

[48] Benyam Kinde, Harrison W. Gabel, Caitlin S. Gilbert, Eric C. Griffith, and Michael E. Greenberg. Reading the unique DNA methylation landscape of the brain: Non-CpG methylation, hydroxymethylation, and MeCP2. *Proceedings of the National Academy of Sciences*, 112(22):6800–6806, June 2015. Publisher: Proceedings of the National Academy of Sciences.

[49] Alexey Kozlenkov, Junhao Li, Pasha Apontes, Yasmin L. Hurd, William M. Byne, Eugene V. Koonin, Michael Wegner, Eran A. Mukamel, and Stella Dracheva. A unique role for DNA (hydroxy)methylation in epigenetic regulation of human inhibitory neurons. *Science Advances*, 4(9):eaau6190, September 2018.

[50] Jana Jeschke, Evelyne Collignon, and François Fuks. Portraits of TET-mediated DNA hydroxymethylation in cancer. *Current Opinion in Genetics & Development*, 36:16–26, February 2016.

[51] Rebecca L. Siegel, Kimberly D. Miller, and Ahmedin Jemal. Cancer statistics, 2019. *CA: A Cancer Journal for Clinicians*, 69(1):7–34, 2019.

[52] David Crosby, Sangeeta Bhatia, Kevin M. Brindle, Lisa M. Coussens, Caroline Dive, Mark Emberton, Sadik Esener, Rebecca C. Fitzgerald, Sanjiv S. Gambhir, Peter Kuhn, Timothy R. Rebbeck, and Shankar Balasubramanian. Early detection of cancer. *Science (New York, N.Y.)*, 375(6586):eaay9040, March 2022.

[53] Yaping Liu. At the dawn: cell-free DNA fragmentomics and gene regulation. *British Journal of Cancer*, 126(3):379–390, February 2022. Number: 3 Publisher: Nature Publishing Group.

[54] Jamie E Medina, Nicholas C Dracopoli, Peter B Bach, Anna Lau, Robert B Scharpf, Gerrit A Meijer, Claus Lindbjerg Andersen, and Victor E Velculescu. Cell-free DNA approaches for cancer early detection and interception. *Journal for Immunotherapy of Cancer*, 11(9):e006013, September 2023.

[55] Arash Jamshidi, Minetta C. Liu, Eric A. Klein, Oliver Venn, Earl Hubbell, John F. Beausang, Samuel Gross, Collin Melton, Alexander P. Fields, Qinwen Liu, Nan Zhang, Eric T. Fung, Kathryn N. Kurtzman, Hamed Amini, Craig Betts, Daniel Civello, Peter Freese, Robert Calef, Konstantin Davydov, Saniya Fayzullina, Chenlu Hou, Roger Jiang, Byoungsok Jung, Susan Tang, Vasiliki Demas, Joshua Newman, Onur Sakarya, Eric Scott, Archana Shenoy, Seyedmehdi Shojaee, Kristan K. Steffen, Virgil Nicula, Tom C. Chien, Siddhartha Bagaria, Nathan Hunkapiller, Mohini Desai, Zhao Dong, Donald A. Richards, Timothy J. Yeatman, Allen L. Cohn, David D. Thiel, Donald A. Berry, Mohan K. Tummala, Kristi McIntyre, Mikkael A. Sekeres, Alan Bryce, Alexander M. Aravanis, Michael V. Seiden, and Charles Swanton. Evaluation of cell-free DNA approaches for multi-cancer early detection. *Cancer Cell*, 40(12):1537–1549.e12, December 2022.

[56] The National Lung Screening Trial Research Team. Reduced Lung-Cancer Mortality with Low-Dose Computed Tomographic Screening. *New England Journal of Medicine*, 365(5):395–409, August 2011.

[57] Xiaolin Wu, Jiahui Li, Asmae Gassa, Denise Buchner, Hakan Alakus, Qiongzhu Dong, Ning Ren, Ming Liu, Margarete Odenthal, Dirk Stippel, Christiane Bruns, Yue Zhao, and Roger Wahba. Circulating tumor DNA as an emerging liquid biopsy biomarker for early diagnosis and therapeutic monitoring in hepatocellular carcinoma. *International Journal of Biological Sciences*, 16(9):1551–1562, 2020.

[58] Peilong Li, Shibiao Liu, Lutao Du, Ghazal Mohseni, Yi Zhang, and Chuanxin Wang. Liquid biopsies based on DNA methylation as biomarkers for the detection and prognosis of lung cancer. *Clinical Epigenetics*, 14(1):118, September 2022.

[59] Stephen Cristiano, Alessandro Leal, Jillian Phallen, Jacob Fiksel, Vilmos Adleff, Daniel C. Bruhm, Sarah Østrup Jensen, Jamie E. Medina, Carolyn Hruban, James R. White, Doreen N. Palsgrove, Noushin Niknafs, Valsamo Anagnostou, Patrick Forde, Jarushka Naidoo, Kristen Marrone, Julie Brahmer, Brian D. Woodward, Hatim Husain, Karlijn L. van Rooijen, Mai-Britt Worm Ørntoft, Anders Husted Madsen, Cornelis J. H. van de Velde, Marcel Verheij, Annemieke Cats, Cornelis J. A. Punt, Geraldine R. Vink, Nicole C. T. van Grieken, Miriam Koopman, Remond J. A. Fijneman, Julia S. Johansen, Hans Jørgen Nielsen, Gerrit A. Meijer, Claus Lindbjerg Andersen, Robert B. Scharpf, and Victor E. Velculescu. Genome-wide cell-free DNA fragmentation in patients with cancer. *Nature*, 570(7761):385–389, June 2019. Number: 7761 Publisher: Nature Publishing Group.

[60] Spencer C. Ding and Y.M. Dennis Lo. Cell-Free DNA Fragmentomics in Liquid Biopsy. *Diagnostics*, 12(4):978, April 2022.

[61] Philip Burnham, Kiran Khush, and Iwijn De Vlaminck. Myriad Applications of Circulating Cell-Free DNA in Precision Organ Transplant Monitoring. *Annals of the American Thoracic Society*, 14(Supplement_3):S237–S241, September 2017.

[62] Anatoli Kustanovich, Ruth Schwartz, Tamar Peretz, and Albert Grinshpun. Life and death of circulating cell-free DNA. *Cancer Biology & Therapy*, 20(8):1057–1067, August 2019.

[63] Ymke van der Pol and Florent Mouliere. Toward the Early Detection of Cancer by Decoding the Epigenetic and Environmental Fingerprints of Cell-Free DNA. *Cancer Cell*, 36(4):350–368, October 2019.

[64] Alexandre Pellan Cheng, Philip Burnham, John Richard Lee, Matthew Pellan Cheng, Manikkam Suthanthiran, Darshana Dadhania, and Iwijn De Vlaminck. A cell-free DNA metagenomic sequencing assay that integrates the host injury response to infection. *Proceedings of the National Academy of Sciences of the United States of America*, 116(37):18738–18744, September 2019.

[65] Benjamin L. Spector, Lauren Harrell, Drinnan Sante, Gerald J. Wyckoff, and Laurel Willig. The methylome and cell-free DNA: current applications in medicine and pediatric disease. *Pediatric Research*, 94(1):89–95, July 2023. Number: 1 Publisher: Nature Publishing Group.

[66] Peiyong Jiang, Kun Sun, Yu K. Tong, Suk Hang Cheng, Timothy H. T. Cheng, Macy M. S. Heung, John Wong, Vincent W. S. Wong, Henry L. Y. Chan, K. C. Allen Chan, Y. M. Dennis Lo, and Rossa W. K. Chiu. Preferred end coordinates and somatic variants as signatures of circulating tumor DNA associated with hepatocellular carcinoma. *Proceedings of the National Academy of Sciences*, 115(46):E10925–E10933, November 2018. Publisher: Proceedings of the National Academy of Sciences.

[67] Kun Sun, Peiyong Jiang, Suk Hang Cheng, Timothy H.T. Cheng, John Wong, Vincent W.S. Wong, Simon S.M. Ng, Brigette B.Y. Ma, Tak Y. Leung, Stephen L. Chan, Tony S.K. Mok, Paul B.S. Lai, Henry L.Y. Chan, Hao Sun, K.C. Allen Chan, Rossa W.K. Chiu, and Y.M. Dennis Lo. Orientation-aware plasma cell-free DNA fragmentation analysis in open chromatin regions informs tissue of origin. *Genome Research*, 29(3):418–427, March 2019.

[68] Alexander Koch, Sophie C. Joosten, Zheng Feng, Tim C. de Ruijter, Muriel X. Draht, Veerle Melotte, Kim M. Smits, Jurgen Veeck, James G. Herman, Leander Van Neste, Wim Van Criekinge, Tim De Meyer, and Manon van Engeland. Analysis of DNA methylation in cancer: location revisited. *Nature Reviews Clinical Oncology*, 15(7):459–466, July 2018. Number: 7 Publisher: Nature Publishing Group.

[69] Stephen B. Baylin, Manel Esteller, Michael R. Rountree, Kurtis E. Bachman, Kornel Schuebel, and James G. Herman. Aberrant patterns of DNA methylation, chromatin formation and gene expression in cancer. *Human Molecular Genetics*, 10(7):687–692, April 2001.

[70] Lili Milani, Anders Lundmark, Anna Kiialainen, Jessica Nordlund, Trond Flaegstad, Erik Forestier, Mats Heyman, Gudmundur Jonmundsson, Jukka Kanerva, Kjeld Schmiegelow, Stefan Söderhäll, Mats G. Gustafsson, Gudmar

Lönnerholm, and Ann-Christine Syvänen. DNA methylation for subtype classification and prediction of treatment outcome in patients with childhood acute lymphoblastic leukemia. *Blood*, 115(6):1214–1225, February 2010.

[71] Saife N. Lone, Sabah Nisar, Tariq Masoodi, Mayank Singh, Arshi Rizwan, Sheema Hashem, Wael El-Rifai, Davide Bedognetti, Surinder K. Batra, Mohammad Haris, Ajaz A. Bhat, and Muzafar A. Macha. Liquid biopsy: a step closer to transform diagnosis, prognosis and future of cancer treatments. *Molecular Cancer*, 21(1):79, March 2022.

[72] M.C. Liu, G.R. Oxnard, E.A. Klein, C. Swanton, M.V. Seiden, Steven R. Cummings, Farnaz Absalan, Gregory Alexander, Brian Allen, Hamed Amini, Alexander M. Aravanis, Siddhartha Bagaria, Leila Bazargan, John F. Beausang, Jennifer Berman, Craig Betts, Alexander Blocker, Joerg Bredno, Robert Calef, Gordon Cann, Jeremy Carter, Christopher Chang, Hemanshi Chawla, Xiaoji Chen, Tom C. Chien, Daniel Civello, Konstantin Davydov, Vasiliki Demas, Mohini Desai, Zhao Dong, Saniya Fayzullina, Alexander P. Fields, Darya Filippova, Peter Freese, Eric T. Fung, Sante Gnerre, Samuel Gross, Meredith Halks-Miller, Megan P. Hall, Anne-Renee Hartman, Chenlu Hou, Earl Hubbell, Nathan Hunkapiller, Karthik Jagadeesh, Arash Jamshidi, Roger Jiang, Byoungsok Jung, TaeHyung Kim, Richard D. Klausner, Kathryn N. Kurtzman, Mark Lee, Wendy Lin, Jafi Lipson, Hai Liu, Qinwen Liu, Margarita Lopatin, Tara Maddala, M. Cyrus Maher, Collin Melton, Andrea Mich, Shivani Nautiyal, Jonathan Newman, Joshua Newman, Virgil Nicula, Cosmos Nicolaou, Ongjen Nikolic, Wenying Pan, Shilpen Patel, Sarah A. Prins, Richard Rava, Neda Ronaghi, Onur Sakarya, Ravi Vijaya Satya, Jan Schellenberger, Eric Scott, Amy J. Sehnert, Rita Shaknovich, Avinash Shanmugam, K.C. Shashidhar, Ling Shen, Archana Shenoy, Seyedmehdi Shojaee, Pranav Singh, Kristan K. Steffen, Susan Tang, Jonathan M. Toung, Anton Valouev, Oliver Venn, Richard T. Williams, Tony Wu, Hui H. Xu, Christopher Yakym, Xiao Yang, Jessica Yecies, Alexander S. Yip, Jack Youngren, Jeanne Yue, Jingyang Zhang, Lily Zhang, Lori (Quan) Zhang, Nan Zhang, Christina Curtis, and Donald A. Berry. Sensitive and specific

multi-cancer detection and localization using methylation signatures in cell-free DNA. *Annals of Oncology*, 31(6):745–759, June 2020.

[73] Peter Ulz, Gerhard G. Thallinger, Martina Auer, Ricarda Graf, Karl Kashofer, Stephan W. Jahn, Luca Abete, Gunda Pristauz, Edgar Petru, Jochen B. Geigl, Ellen Heitzer, and Michael R. Speicher. Inferring expressed genes by whole-genome sequencing of plasma DNA. *Nature Genetics*, 48(10):1273–1278, October 2016.

[74] Matthew W. Snyder, Martin Kircher, Andrew J. Hill, Riza M. Daza, and Jay Shendure. Cell-free DNA Comprises an In Vivo Nucleosome Footprint that Informs Its Tissues-Of-Origin. *Cell*, 164(1-2):57–68, January 2016.

[75] Mohammad Shahrokh Esfahani, Emily G. Hamilton, Mahya Mehrmohamadi, Barzin Y. Nabet, Stefan K. Alig, Daniel A. King, Chloé B. Steen, Charles W. Macaulay, Andre Schultz, Monica C. Nesselbush, Joanne Soo, Joseph G. Schroers-Martin, Binbin Chen, Michael S. Binkley, Henning Stehr, Jacob J. Chabon, Brian J. Sworder, Angela B.-Y. Hui, Matthew J. Frank, Everett J. Moding, Chih Long Liu, Aaron M. Newman, James M. Isbell, Charles M. Rudin, Bob T. Li, David M. Kurtz, Maximilian Diehn, and Ash A. Alizadeh. Inferring gene expression from cell-free DNA fragmentation profiles. *Nature Biotechnology*, 40(4):585–597, April 2022. Number: 4 Publisher: Nature Publishing Group.

[76] M. Cisneros-Villanueva, L. Hidalgo-Pérez, M. Rios-Romero, A. Cedro-Tanda, C. A. Ruiz-Villavicencio, K. Page, R. Hastings, D. Fernandez-Garcia, R. Allsopp, M. A. Fonseca-Montaño, S. Jimenez-Morales, V. Padilla-Palma, J. A. Shaw, and A. Hidalgo-Miranda. Cell-free DNA analysis in current cancer clinical trials: a review. *British Journal of Cancer*, 126(3):391–400, February 2022. Number: 3 Publisher: Nature Publishing Group.

[77] T F Greten, F Papendorf, J S Bleck, T Kirchhoff, T Wohlberedt, S Kubicka, J Klempnauer, M Galanski, and M P Manns. Survival rate in patients with hepatocellular carcinoma: a retrospective analysis of 389 patients. *British Journal of Cancer*, 92(10):1862–1868, May 2005.

[78] Alexander P. Stark, Greg D. Sacks, Matthew M. Rochefort, Timothy R. Donahue, Howard A. Reber, James S. Tomlinson, David W. Dawson, Guido Eibl, and O. Joe Hines. Long-term Survival in Patients with Pancreatic Ductal Adenocarcinoma. *Surgery*, 159(6):1520–1527, June 2016.

[79] Sonia Pascual, Cayetano Miralles, Juan M Bernabé, Javier Irurzun, and Mariana Planells. Surveillance and diagnosis of hepatocellular carcinoma: A systematic review. *World Journal of Clinical Cases*, 7(16):2269–2286, August 2019.

[80] F. Trevisani, P. E. D'Intino, A. M. Morselli-Labate, G. Mazzella, E. Accogli, P. Caraceni, M. Domenicali, S. De Notariis, E. Roda, and M. Bernardi. Serum alpha-fetoprotein for diagnosis of hepatocellular carcinoma in patients with chronic liver disease: influence of HBsAg and anti-HCV status. *Journal of Hepatology*, 34(4):570–575, April 2001.

[81] Saowanee Ngamruengphong and Marcia Irene Canto. Screening for Pancreatic Cancer. *The Surgical clinics of North America*, 96(6):1223–1233, December 2016.

[82] K. C. Allen Chan, Peiyong Jiang, Carol W. M. Chan, Kun Sun, John Wong, Edwin P. Hui, Stephen L. Chan, Wing Cheong Chan, David S. C. Hui, Simon S. M. Ng, Henry L. Y. Chan, Cesar S. C. Wong, Brigette B. Y. Ma, Anthony T. C. Chan, Paul B. S. Lai, Hao Sun, Rossa W. K. Chiu, and Y. M. Dennis Lo. Noninvasive detection of cancer-associated genome-wide hypomethylation and copy number aberrations by plasma DNA bisulfite sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, 110(47):18761–18768, November 2013.

[83] Kun Sun, Peiyong Jiang, K. C. Allen Chan, John Wong, Yvonne K. Y. Cheng, Raymond H. S. Liang, Wai-kong Chan, Edmond S. K. Ma, Stephen L. Chan, Suk Hang Cheng, Rebecca W. Y. Chan, Yu K. Tong, Simon S. M. Ng, Raymond S. M. Wong, David S. C. Hui, Tse Ngong Leung, Tak Y. Leung, Paul B. S. Lai, Rossa W. K. Chiu, and Yuk Ming Dennis Lo. Plasma DNA

tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments. *Proceedings of the National Academy of Sciences*, 112(40):E5503–E5512, October 2015. Publisher: Proceedings of the National Academy of Sciences.

[84] Shuli Kang, Qingjiao Li, Quan Chen, Yonggang Zhou, Stacy Park, Gina Lee, Brandon Grimes, Kostyantyn Krysan, Min Yu, Wei Wang, Frank Alber, Fengzhu Sun, Steven M. Dubinett, Wenyuan Li, and Xianghong Jasmine Zhou. CancerLocator: non-invasive cancer diagnosis and tissue-of-origin prediction using methylation profiles of cell-free DNA. *Genome Biology*, 18(1):53, December 2017.

[85] Joshua Moss, Judith Magenheim, Daniel Neiman, Hai Zemmour, Netanel Loyfer, Amit Korach, Yaacov Samet, Myriam Maoz, Henrik Druid, Peter Arner, Keng-Yeh Fu, Endre Kiss, Kirsty L. Spalding, Giora Landesberg, Aviad Zick, Albert Grinshpun, A. M. James Shapiro, Markus Grompe, Avigail Dreazan Wittenberg, Benjamin Glaser, Ruth Shemer, Tommy Kaplan, and Yuval Dor. Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease. *Nature Communications*, 9:5068, November 2018.

[86] C. S. Wilhelm-Benartzi, D. C. Koestler, M. R. Karagas, J. M. Flanagan, B. C. Christensen, K. T. Kelsey, C. J. Marsit, E. A. Houseman, and R. Brown. Review of processing and analysis methods for DNA methylation array data. *British Journal of Cancer*, 109(6):1394–1402, September 2013. Number: 6 Publisher: Nature Publishing Group.

[87] Ruth Pidsley, Elena Zotenko, Timothy J. Peters, Mitchell G. Lawrence, Gail P. Risbridger, Peter Molloy, Susan Van Djik, Beverly Muhlhausler, Clare Stirzaker, and Susan J. Clark. Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biology*, 17:208, October 2016.

[88] Shicai Fan, Chengzhe Li, Rizi Ai, Mengchi Wang, Gary S. Firestein, and Wei Wang. Computationally expanding infinium HumanMethylation450 BeadChip array data to reveal distinct DNA methylation patterns of rheumatoid arthritis. *Bioinformatics*, 32(12):1773–1778, June 2016.

[89] Yuanyuan Li and Trygve O. Tollefsbol. DNA methylation detection: Bisulfite genomic sequencing analysis. *Methods in molecular biology (Clifton, N.J.)*, 791:11–21, 2011.

[90] Minetta C. Liu, Arash Jamshidi, Oliver Venn, Alexander P. Fields, M. Cyrus Maher, Gordon Cann, Hamed Amini, Samuel Gross, Joerg Bredno, Meredith Miller, Jan Schellenberger, Kathryn N. Kurtzman, Eric T. Fung, Tara Maddala, Geoffrey R. Oxnard, Eric A. Klein, David R. Spigel, Anne-Renee Hartman, Alex Aravanis, and Michael Seiden. Genome-wide cell-free DNA (cfDNA) methylation signatures and effect on tissue of origin (TOO) performance. *Journal of Clinical Oncology*, 37(15_suppl):3049–3049, May 2019. Publisher: Wolters Kluwer.

[91] Netanel Loyfer, Judith Magenheim, Ayelet Peretz, Gordon Cann, Joerg Bredno, Agnes Klochendler, Ilana Fox-Fisher, Sapir Shabi-Porat, Merav Hecht, Tsuria Pelet, Joshua Moss, Zeina Drawshy, Hamed Amini, Patriss Moradi, Sudharani Nagaraju, Dvora Bauman, David Shveiky, Shay Porat, Uri Dior, Gurion Rivkin, Omer Or, Nir Hirshoren, Einat Carmon, Alon Pikarsky, Abed Khalaileh, Gideon Zamir, Ronit Grinbaum, Machmud Abu Gazala, Ido Mizrahi, Noam Shussman, Amit Korach, Ori Wald, Uzi Izhar, Eldad Erez, Vladimir Yutkin, Yaacov Samet, Devorah Rotnemer Golinkin, Kirsty L. Spalding, Henrik Druid, Peter Arner, A. M. James Shapiro, Markus Grompe, Alex Aravanis, Oliver Venn, Arash Jamshidi, Ruth Shemer, Yuval Dor, Benjamin Glaser, and Tommy Kaplan. A DNA methylation atlas of normal human cell types. *Nature*, 613(7943):355–364, January 2023. Number: 7943 Publisher: Nature Publishing Group.

[92] Kazuo Tanaka and Akimitsu Okamoto. Degradation of DNA by bisulfite treatment. *Bioorganic & Medicinal Chemistry Letters*, 17(7):1912–1915, April 2007.

[93] Yibin Liu, Paulina Siejka-Zielińska, Gergana Velikova, Ying Bi, Fang Yuan, Marketa Tomkova, Chunsen Bai, Lei Chen, Benjamin Schuster-Böckler, and Chun-Xiao Song. Bisulfite-free direct detection of 5-methylcytosine and 5-hydroxymethylcytosine at base resolution. *Nature Biotechnology*, 37(4):424–429, April 2019. Number: 4 Publisher: Nature Publishing Group.

[94] Anshul Kundaje, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, Pouya Kheradpour, Zhizhuo Zhang, Jianrong Wang, Michael J. Ziller, Viren Amin, John W. Whitaker, Matthew D. Schultz, Lucas D. Ward, Abhishek Sarkar, Gerald Quon, Richard S. Sandstrom, Matthew L. Eaton, Yi-Chieh Wu, Andreas R. Pfenning, Xinchen Wang, Melina Claussnitzer, Yaping Liu, Cristian Coarfa, R. Alan Harris, Noam Shoresh, Charles B. Epstein, Elizabeta Gjoneska, Danny Leung, Wei Xie, R. David Hawkins, Ryan Lister, Chibo Hong, Philippe Gascard, Andrew J. Mungall, Richard Moore, Eric Chuah, Angela Tam, Theresa K. Canfield, R. Scott Hansen, Rajinder Kaul, Peter J. Sabo, Mukul S. Bansal, Annaick Carles, Jesse R. Dixon, Kai-How Farh, Soheil Feizi, Rosa Karlic, Ah-Ram Kim, Ashwinikumar Kulkarni, Daofeng Li, Rebecca Lowdon, GiNell Elliott, Tim R. Mercer, Shane J. Neph, Vitor Onuchic, Paz Polak, Nisha Rajagopal, Pradipta Ray, Richard C. Sallari, Kyle T. Siebenthall, Nicholas A. Sinnott-Armstrong, Michael Stevens, Robert E. Thurman, Jie Wu, Bo Zhang, Xin Zhou, Arthur E. Beaudet, Laurie A. Boyer, Philip L. De Jager, Peggy J. Farnham, Susan J. Fisher, David Haussler, Steven J. M. Jones, Wei Li, Marco A. Marra, Michael T. McManus, Shamil Sunyaev, James A. Thomson, Thea D. Tlsty, Li-Huei Tsai, Wei Wang, Robert A. Waterland, Michael Q. Zhang, Lisa H. Chadwick, Bradley E. Bernstein, Joseph F. Costello, Joseph R. Ecker, Martin Hirst, Alexander Meissner, Aleksandar Milosavljevic, Bing Ren, John A. Stamatoyannopoulos, Ting Wang, and Manolis Kellis. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330, February 2015. Number: 7539 Publisher: Nature Publishing Group.

[95] Mark D. Robinson, Abdullah Kahraman, Charity W. Law, Helen Lindsay, Malgorzata Nowicka, Lukas M. Weber, and Xiaobei Zhou. Statistical meth-

ods for detecting differentially methylated loci and regions. *Frontiers in Genetics*, 5, September 2014.

[96] Ding-Xuan Zhou. Universality of deep convolutional neural networks. *Applied and Computational Harmonic Analysis*, 48(2):787–794, March 2020.

[97] B. van Ginneken, A. A. A. Setio, C. Jacobs, and F. Ciompi. Off-the-shelf convolutional neural network features for pulmonary nodule detection in computed tomography scans. In *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, pages 286–289, April 2015.

[98] Yikuan Li, Shishir Rao, Jose Roberto Ayala Solares, Abdelaali Hassaine, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. BEHRT: Transformer for Electronic Health Records. *arXiv:1907.09538 [cs, stat]*, July 2019.

[99] Christof Angermueller, Heather J. Lee, Wolf Reik, and Oliver Stegle. DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. *Genome Biology*, 18(1):67, December 2017.

[100] Qi Tian, Jianxiao Zou, Jianxiong Tang, Yuan Fang, Zhongli Yu, and Shicai Fan. MRCNN: a deep learning model for regression of genome-wide DNA methylation. *BMC Genomics*, 20(S2):192, April 2019.

[101] Jiaqi Li, Lei Wei, Xianglin Zhang, Wei Zhang, Haochen Wang, Bixi Zhong, Zhen Xie, Hairong Lv, and Xiaowo Wang. DISMIR: Deep learning-based noninvasive cancer detection by integrating DNA sequence and methylation information of individual cell-free DNA reads. *Briefings in Bioinformatics*, 22(6):bbab250, November 2021.

[102] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David

Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, August 2021. Number: 7873 Publisher: Nature Publishing Group.

[103] Daniel Quang and Xiaohui Xie. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Research*, 44(11):e107, June 2016.

[104] Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R. Ledsam, Agnieszka Grabska-Barwinska, Kyle R. Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R. Kelley. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods*, 18(10):1196–1203, October 2021. Number: 10 Publisher: Nature Publishing Group.

[105] Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. *arXiv:1409.1259 [cs, stat]*, October 2014.

[106] James Martens and Ilya Sutskever. Learning Recurrent Neural Networks with Hessian-Free Optimization. page 8.

[107] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. *arXiv:1706.03762 [cs]*, December 2017.

[108] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*, May 2019.

[109] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey

Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners. *arXiv:2005.14165 [cs]*, July 2020.

[110] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, March 2023. Publisher: American Association for the Advancement of Science.

[111] Ali Madani, Ben Krause, Eric R. Greene, Subu Subramanian, Benjamin P. Mohr, James M. Holton, Jose Luis Olmos, Caiming Xiong, Zachary Z. Sun, Richard Socher, James S. Fraser, and Nikhil Naik. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, 41(8):1099–1106, August 2023. Number: 8 Publisher: Nature Publishing Group.

[112] Johannes Linder, Divyanshi Srivastava, Han Yuan, Vikram Agarwal, and David R. Kelley. Predicting RNA-seq coverage from DNA sequence as a unifying model of gene regulation. preprint, Genomics, September 2023.

[113] Eric Nguyen, Michael Poli, Marjan Faizi, Armin Thomas, Callum Birch-Sykes, Michael Wornow, Aman Patel, Clayton Rabideau, Stefano Massaroli, Yoshua Bengio, Stefano Ermon, Stephen A. Baccus, and Chris Ré. HyenaDNA: Long-Range Genomic Sequence Modeling at Single Nucleotide Resolution, November 2023. arXiv:2306.15794 [cs, q-bio].

[114] Micaela E. Consens, Cameron Dufault, Michael Wainberg, Duncan Forster, Mehran Karimzadeh, Hani Goodarzi, Fabian J. Theis, Alan Moses, and Bo Wang. To Transformers and Beyond: Large Language Models for the Genome, November 2023. arXiv:2311.07621 [cs, q-bio].

[115] Kasper D Hansen, Benjamin Langmead, and Rafael A Irizarry. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biology*, 13(10):R83, 2012.

[116] Egor Dolzhenko and Andrew D. Smith. Using beta-binomial regression for high-precision differential methylation analysis in multifactor whole-genome bisulfite sequencing experiments. *BMC Bioinformatics*, 15(1):215, June 2014.

[117] Yongseok Park, Maria E. Figueroa, Laura S. Rozek, and Maureen A. Sartor. MethylSig: a whole genome DNA methylation analysis pipeline. *Bioinformatics (Oxford, England)*, 30(17):2414–2422, September 2014.

[118] Frank Jühling, Helene Kretzmer, Stephan H. Bernhart, Christian Otto, Peter F. Stadler, and Steve Hoffmann. metilene: fast and sensitive calling of differentially methylated regions from bisulfite sequencing data. *Genome Research*, 26(2):256–262, February 2016.

[119] John M. Gaspar and Ronald P. Hart. DMRfinder: efficiently identifying differentially methylated regions from MethylC-seq data. *BMC Bioinformatics*, 18(1):528, November 2017.

[120] Timothy J Peters, Michael J Buckley, Yunshun Chen, Gordon K Smyth, Christopher C Goodnow, and Susan J Clark. Calling differentially methylated regions from whole genome bisulphite sequencing with DMRcate. *Nucleic Acids Research*, 49(19):e109, November 2021.

[121] Christa Caggiano, Barbara Celona, Fleur Garton, Joel Mefford, Brian L. Black, Robert Henderson, Catherine Lomen-Hoerth, Andrew Dahl, and Noah Zaitlen. Comprehensive cell type decomposition of circulating cell-free DNA with CelFiE. *Nature Communications*, 12(1):2717, May 2021. Number: 1 Publisher: Nature Publishing Group.

[122] Yalu Wen, Fushun Chen, Qingzheng Zhang, Yan Zhuang, and Zhiguang Li. Detection of differentially methylated regions in whole genome bisulfite sequencing data using local Getis-Ord statistics. *Bioinformatics*, page btw497, August 2016.

[123] Adib Shafi, Cristina Mitrea, Tin Nguyen, and Sorin Draghici. A survey of the approaches for identifying differential methylation using bisulfite sequencing data. *Briefings in Bioinformatics*, 19(5):737–753, September 2018.

[124] Wenyuan Li, Qingjiao Li, Shuli Kang, Mary Same, Yonggang Zhou, Carol Sun, Chun-Chi Liu, Lea Matsuoka, Linda Sher, Wing Hung Wong, Frank Alber, and Xianghong Jasmine Zhou. CancerDetector: ultrasensitive and non-invasive cancer detection at the resolution of individual reads using cell-free DNA methylation sequencing data. *Nucleic Acids Research*, 46(15):e89–e89, September 2018.

[125] Eric Chater-Diehl, Sarah J. Goodman, Cheryl Cytrynbaum, Andrei L. Turinsky, Sanaa Choufani, and Rosanna Weksberg. Anatomy of DNA methylation signatures: Emerging insights and applications. *The American Journal of Human Genetics*, 108(8):1359–1366, August 2021.

[126] Steve Horvath. DNA methylation age of human tissues and cell types. *Genome Biology*, 14(10):3156, December 2013.

[127] Rebecca L. Hood, Laila C. Schenkel, Sarah M. Nikkel, Peter J. Ainsworth, Guillaume Pare, Kym M. Boycott, Dennis E. Bulman, and Bekim Sadikovic. The defining DNA methylation signature of Floating-Harbor Syndrome. *Scientific Reports*, 6:38803, December 2016.

[128] Matthias Wielscher, Pooja R. Mandaviya, Brigitte Kuehnel, Roby Joehanes, Rima Mustafa, Oliver Robinson, Yan Zhang, Barbara Bodinier, Esther Walton, Pashupati P. Mishra, Pascal Schlosser, Rory Wilson, Pei-Chien Tsai, Saranya Palaniswamy, Riccardo E. Marioni, Giovanni Fiorito, Giovanni Cugliari, Ville Karhunen, Mohsen Ghanbari, Bruce M. Psaty, Marie Loh, Joshua C. Bis, Benjamin Lehne, Nona Sotoodehnia, Ian J. Deary, Marc Chadeau-Hyam, Jennifer A. Brody, Alexia Cardona, Elizabeth Selvin, Alicia K. Smith, Andrew H. Miller, Mylin A. Torres, Eirini Marouli, Xin Gào, Joyce B. J. van Meurs, Johanna Graf-Schindler, Wolfgang Rathmann, Wolfgang Koenig, Annette Peters, Wolfgang Weninger, Matthias Farlik, Tao Zhang, Wei Chen, Yujing Xia, Alexander Teumer, Matthias Nauck, Hans J.

Grabe, Macus Doerr, Terho Lehtimäki, Weihua Guan, Lili Milani, Toshiko Tanaka, Krista Fisher, Lindsay L. Waite, Silva Kasela, Paolo Vineis, Niek Verweij, Pim van der Harst, Licia Iacoviello, Carlotta Sacerdote, Salvatore Panico, Vittorio Krogh, Rosario Tumino, Evangelia Tzala, Giuseppe Matullo, Mikko A. Hurme, Olli T. Raitakari, Elena Colicino, Andrea A. Baccarelli, Mika Kähönen, Karl-Heinz Herzig, Shengxu Li, Karen N. Conneely, Jaspal S. Kooner, Anna Köttgen, Bastiaan T. Heijmans, Panos Deloukas, Caroline Relton, Ken K. Ong, Jordana T. Bell, Eric Boerwinkle, Paul Elliott, Hermann Brenner, Marian Beekman, Daniel Levy, Melanie Waldenberger, John C. Chambers, Abbas Dehghan, and Marjo-Riitta Järvelin. DNA methylation signature of chronic low-grade inflammation and its role in cardio-respiratory diseases. *Nature Communications*, 13(1):2408, May 2022. Number: 1 Publisher: Nature Publishing Group.

[129] Christine Ladd-Acosta and M. Daniele Fallin. DNA methylation signatures as biomarkers of prior environmental exposures. *Current epidemiology reports*, 6(1):1–13, March 2019.

[130] Philipp Jurmeister, Michael Bockmayr, Philipp Seegerer, Teresa Bockmayr, Denise Treue, Grégoire Montavon, Claudia Vollbrecht, Alexander Arnold, Daniel Teichmann, Keno Bressem, Ulrich Schüller, Maximilian von Laffert, Klaus-Robert Müller, David Capper, and Frederick Klauschen. Machine learning analysis of DNA methylation profiles distinguishes primary lung squamous cell carcinomas from head and neck metastases. *Science Translational Medicine*, 11(509):eaaw8513, September 2019. Publisher: American Association for the Advancement of Science.

[131] Steve O'Hagan, Marina Wright Muelas, Philip J. Day, Emma Lundberg, and Douglas B. Kell. GeneGini: Assessment via the Gini Coefficient of Reference "Housekeeping" Genes and Diverse Human Transporter Expression Profiles. *Cell Systems*, 6(2):230–244.e1, February 2018.

[132] Jonathan Schug, Winfried-Paul Schuller, Claudia Kappen, J. Michael Salbaum, Maja Bucan, and Christian J. Stoeckert. Promoter features related

to tissue specificity as measured by Shannon entropy. *Genome Biology*, 6(4):R33, March 2005.

[133] Andrzej Dudek. Silhouette Index as Clustering Evaluation Tool. In Krzysztof Jajuga, Jacek Batóg, and Marek Walesiak, editors, *Classification and Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization, pages 19–33, Cham, 2020. Springer International Publishing.

[134] José E. Chacón and Ana I. Rastrojo. Minimum adjusted Rand index for two clusterings of a given size. *Advances in Data Analysis and Classification*, 17(1):125–133, March 2023.

[135] Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance. *Journal of Machine Learning Research*, 11(95):2837–2854, 2010.

[136] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3(null):993–1022, March 2003.

[137] Kevin Menden, Mohamed Marouf, Sergio Oller, Anupriya Dalmia, Daniel Sumner Magruder, Karin Kloiber, Peter Heutink, and Stefan Bonn. Deep learning–based cell composition analysis from tissue expression profiles. *Science Advances*, 6(30):eaba2619, July 2020. Publisher: American Association for the Advancement of Science.

[138] Yanshuo Chen, Yixuan Wang, Yuelong Chen, Yuqi Cheng, Yumeng Wei, Yunxiang Li, Jiuming Wang, Yingying Wei, Ting-Fung Chan, and Yu Li. Deep autoencoder for interpretable tissue-adaptive deconvolution and cell-type-specific gene analysis. *Nature Communications*, 13(1):6735, November 2022. Number: 1 Publisher: Nature Publishing Group.

[139] Pia Keukeleire, Stavros Makrodimitris, and Marcel Reinders. Cell type deconvolution of methylated cell-free DNA at the resolution of individual reads. *NAR Genomics and Bioinformatics*, 5(2):lqad048, June 2023.

[140] E. Andres Houseman, Molly L. Kile, David C. Christiani, Tan A. Ince, Karl T. Kelsey, and Carmen J. Marsit. Reference-free deconvolution of DNA methylation data and mediation by cell composition effects. *BMC Bioinformatics*, 17(1):259, June 2016.

[141] Pavlo Lutsik, Martin Slawski, Gilles Gasparoni, Nikita Vedeneev, Matthias Hein, and Jörn Walter. MeDeCom: discovery and quantification of latent components of heterogeneous methylomes. *Genome Biology*, 18(1):55, March 2017.

[142] Michael Scherer, Petr V. Nazarov, Reka Toth, Shashwat Sahay, Tony Kaoma, Valentin Maurer, Nikita Vedeneev, Christoph Plass, Thomas Lengauer, Jörn Walter, and Pavlo Lutsik. Reference-free deconvolution, visualization and interpretation of complex DNA methylation data using DecompPipeline, MeDeCom and FactorViz. *Nature Protocols*, 15(10):3240–3263, October 2020. Number: 10 Publisher: Nature Publishing Group.

[143] W. James Kent, Charles W. Sugnet, Terrence S. Furey, Krishna M. Roskin, Tom H. Pringle, Alan M. Zahler, and and David Haussler. The Human Genome Browser at UCSC. *Genome Research*, 12(6):996–1006, June 2002.

[144] Paulina Siejka-Zielińska, Jingfei Cheng, Felix Jackson, Yibin Liu, Zahir Soonawalla, Srikanth Reddy, Michael Silva, Luminita Puta, Misti Vanette McCain, Emma L. Culver, Noor Bekkali, Benjamin Schuster-Böckler, Pier Francesco Palamara, Derek Mann, Helen Reeves, Eleanor Barnes, Shivan Sivakumar, and Chun-Xiao Song. Cell-free DNA TAPS provides multimodal information for early cancer detection. *Science Advances*, 7(36):eabh0534, September 2021.

[145] Daniel R. Zerbino, Steven P. Wilder, Nathan Johnson, Thomas Juettemann, and Paul R. Flicek. The Ensembl Regulatory Build. *Genome Biology*, 16(1):56, March 2015.

[146] Je Yeong Ko, Sumin Oh, and Kyung Hyun Yoo. Functional Enhancers As Master Regulators of Tissue-Specific Gene Regulation and Cancer Development. *Molecules and Cells*, 40(3):169–177, March 2017.

[147] Tomás Pachano, Endika Haro, and Alvaro Rada-Iglesias. Enhancer-gene specificity in development and disease. *Development (Cambridge, England)*, 149(11):dev186536, June 2022.

[148] Amandine Barral and Jérôme Déjardin. The chromatin signatures of enhancers and their dynamic regulation. *Nucleus*, 14(1):2160551, December 2023. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/19491034.2022.2160551.

[149] Xueyan Wang, Danli Liu, Jing Luo, Dashuai Kong, and Yubo Zhang. Exploring the Role of Enhancer-Mediated Transcriptional Regulation in Precision Biology. *International Journal of Molecular Sciences*, 24(13):10843, June 2023.

[150] Laurens van der Maaten and Geoffrey Hinton. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

[151] Yongseok Park and Hao Wu. Differential methylation analysis for BS-seq data under general experimental design. *Bioinformatics*, 32(10):1446–1453, May 2016.

[152] Ting Qi, Min Pan, Huajuan Shi, Liangying Wang, Yunfei Bai, and Qinyu Ge. Cell-Free DNA Fragmentomics: The Novel Promising Biomarker. *International Journal of Molecular Sciences*, 24(2):1503, January 2023.

[153] Anne Tranberg Madsen, Johanne Andersen Hojbjerg, Boe Sandahl Sorensen, and Anne Winther-Larsen. Day-to-day and within-day biological variation of cell-free DNA. *EBioMedicine*, 49:284–290, October 2019.

[154] Devin C. Koestler, Joseph Usset, Brock C. Christensen, Carmen J. Marsit, Margaret R. Karagas, Karl T. Kelsey, and John K. Wiencke. DNA methylation-derived neutrophil-to-lymphocyte ratio: an epigenetic tool to explore cancer inflammation and outcomes. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology*, 26(3):328–338, March 2017.

[155] Shuo Li, Weihua Zeng, Xiaohui Ni, Qiao Liu, Wenyuan Li, Mary L. Stackpole, Yonggang Zhou, Arjan Gower, Kostyantyn Krysan, Preeti Ahuja, David S. Lu, Steven S. Raman, William Hsu, Denise R. Aberle, Clara E. Magyar, Samuel W. French, Steven-Huy B. Han, Edward B. Garon, Vatche G. Agopian, Wing Hung Wong, Steven M. Dubinett, and Xianghong Jasmine Zhou. Comprehensive tissue deconvolution of cell-free DNA by deep learning for disease diagnosis and monitoring. *Proceedings of the National Academy of Sciences*, 120(28):e2305236120, July 2023. Company: National Academy of Sciences Distributor: National Academy of Sciences Institution: National Academy of Sciences Label: National Academy of Sciences Publisher: Proceedings of the National Academy of Sciences.

[156] Anastasios N. Angelopoulos and Stephen Bates. A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification, December 2022. arXiv:2107.07511 [cs, math, stat].

[157] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, June 2009. ISSN: 1063-6919.

[158] Andriy Kryshtafovych, Torsten Schwede, Maya Topf, Krzysztof Fidelis, and John Moult. Critical Assessment of Methods of Protein Structure Prediction (CASP) – Round XIII. *Proteins*, 87(12):1011–1020, December 2019.

[159] Florian Eckhardt, Joern Lewin, Rene Cortese, Vardhman K. Rakyan, John Attwood, Matthias Burger, John Burton, Tony V. Cox, Rob Davies, Thomas A. Down, Carolina Haefliger, Roger Horton, Kevin Howe, David K. Jackson, Jan Kunde, Christoph Koenig, Jennifer Liddle, David Niblett, Thomas Otto, Roger Pettett, Stefanie Seemann, Christian Thompson, Tony West, Jane Rogers, Alex Olek, Kurt Berlin, and Stephan Beck. DNA methylation profiling of human chromosomes 6, 20 and 22. *Nature Genetics*, 38(12):1378–1385, December 2006. Number: 12 Publisher: Nature Publishing Group.

[160] Shicheng Guo, Dinh Diep, Nongluk Plongthongkum, Ho-Lim Fung, Kang Zhang, and Kun Zhang. 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):635–642, April 2017. Number: 4 Publisher: Nature Publishing Group.

[161] Jianghan Qu, Meng Zhou, Qiang Song, Elizabeth E. Hong, and Andrew D. Smith. MLML: consistent simultaneous estimates of DNA methylation and hydroxymethylation. *Bioinformatics*, 29(20):2645–2646, October 2013.

[162] Edward Y Chen, Christopher M Tan, Yan Kou, Qiaonan Duan, Zichen Wang, Gabriela Vaz Meirelles, Neil R Clark, and Avi Ma'ayan. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics*, 14:128, April 2013.

[163] John Lonsdale, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, Gary Walters, Fernando Garcia, Nancy Young, Barbara Foster, Mike Moser, Ellen Karasik, Bryan Gillard, Kimberley Ramsey, Susan Sullivan, Jason Bridge, Harold Magazine, John Syron, Johnelle Fleming, Laura Siminoff, Heather Traino, Maghboeba Mosavel, Laura Barker, Scott Jewell, Dan Rohrer, Dan Maxim, Dana Filkins, Philip Harbach, Eddie Cortadillo, Bree Berghuis, Lisa Turner, Eric Hudson, Kristin Feenstra, Leslie Sobin, James Robb, Phillip Branton, Greg Korzeniewski, Charles Shive, David Tabor, Liqun Qi, Kevin Groch, Sreenath Nampally, Steve Buia, Angela Zimmerman, Anna Smith, Robin Burges, Karna Robinson, Kim Valentino, Deborah Bradbury, Mark Cosentino, Norma Diaz-Mayoral, Mary Kennedy, Theresa Engel, Penelope Williams, Kenyon Erickson, Kristin Ardlie, Wendy Winckler, Gad Getz, David DeLuca, Daniel MacArthur, Manolis Kellis, Alexander Thomson, Taylor Young, Ellen Gelfand, Molly Donovan, Yan Meng, George Grant, Deborah Mash, Yvonne Marcus, Margaret Basile, Jun Liu, Jun Zhu, Zhidong Tu, Nancy J. Cox, Dan L. Nicolae, Eric R. Gamazon, Hae Kyung Im, Anuar Konkashbaev, Jonathan Pritchard, Matthew Stevens, Timothèe Flutre, Xiaoquan Wen,

Emmanouil T. Dermitzakis, Tuuli Lappalainen, Roderic Guigo, Jean Monlong, Michael Sammeth, Daphne Koller, Alexis Battle, Sara Mostafavi, Mark McCarthy, Manual Rivas, Julian Maller, Ivan Rusyn, Andrew Nobel, Fred Wright, Andrey Shabalin, Mike Feolo, Nataliya Sharopova, Anne Sturcke, Justin Paschal, James M. Anderson, Elizabeth L. Wilder, Leslie K. Derr, Eric D. Green, Jeffery P. Struewing, Gary Temple, Simona Volpi, Joy T. Boyer, Elizabeth J. Thomson, Mark S. Guyer, Cathy Ng, Assya Abdallah, Deborah Colantuoni, Thomas R. Insel, Susan E. Koester, A. Roger Little, Patrick K. Bender, Thomas Lehner, Yin Yao, Carolyn C. Compton, Jimmie B. Vaught, Sherilyn Sawyer, Nicole C. Lockhart, Joanne Demchok, and Helen F. Moore. The Genotype-Tissue Expression (GTEx) project. *Nature Genetics*, 45(6):580–585, June 2013. Number: 6 Publisher: Nature Publishing Group.

[164] Oscar Franzén, Li-Ming Gan, and Johan L M Björkegren. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. *Database*, 2019:baz046, January 2019.

[165] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, May 2012.

[166] Zhao Li, Bo Ye, Minwei Bao, Binbin Xu, Qinyi Chen, Sida Liu, Yudong Han, Mingzhen Peng, Zhifeng Lin, Jingpei Li, Wenzhuo Zhu, Qiang Lin, and Liwen Xiong. Radiologic Predictors for Clinical Stage IA Lung Adenocarcinoma with Ground Glass Components: A Multi-Center Study of Long-Term Outcomes. *PLOS ONE*, 10(9):e0136616, September 2015.

[167] Maia H. Malonzo and Harri Lähdesmäki. LuxHMM: DNA methylation analysis with genome segmentation via hidden Markov model. *BMC Bioinformatics*, 24:58, February 2023.

[168] Xiao-Long Cui, Ji Nie, Jeremy Ku, Urszula Dougherty, Diana C. West-Szymanski, Francois Collin, Christopher K. Ellison, Laura Sieh, Yuhong Ning, Zifeng Deng, Carolyn W. T. Zhao, Anna Bergamaschi, Joel Pekow,

Jiangbo Wei, Alana V. Beadell, Zhou Zhang, Geeta Sharma, Raman Talwar, Patrick Arensdorf, Jason Karpus, Ajay Goel, Marc Bissonnette, Wei Zhang, Samuel Levy, and Chuan He. A human tissue map of 5-hydroxymethylcytosines exhibits tissue specificity through gene and enhancer modulation. *Nature Communications*, 11(1):6161, December 2020. Number: 1 Publisher: Nature Publishing Group.

[169] Jianfeng Xu, Jiejun Shi, Xiaodong Cui, Ya Cui, Jingyi Jessica Li, Ajay Goel, Xi Chen, Jean-Pierre Issa, Jianzhong Su, and Wei Li. Cellular Heterogeneity–Adjusted cLonal Methylation (CHALM) improves prediction of gene expression. *Nature Communications*, 12(1):400, January 2021. Number: 1 Publisher: Nature Publishing Group.

[170] Mathias Uhlén, Linn Fagerberg, Björn M. Hallström, Cecilia Lindskog, Per Oksvold, Adil Mardinoglu, Åsa Sivertsson, Caroline Kampf, Evelina Sjöstedt, Anna Asplund, IngMarie Olsson, Karolina Edlund, Emma Lundberg, Sanjay Navani, Cristina Al-Khalili Szigyarto, Jacob Odeberg, Dijana Djureinovic, Jenny Ottosson Takanen, Sophia Hober, Tove Alm, Per-Henrik Edqvist, Holger Berling, Hanna Tegel, Jan Mulder, Johan Rockberg, Peter Nilsson, Jochen M. Schwenk, Marica Hamsten, Kalle von Feilitzen, Mattias Forsberg, Lukas Persson, Fredric Johansson, Martin Zwahlen, Gunnar von Heijne, Jens Nielsen, and Fredrik Pontén. Tissue-based map of the human proteome. *Science*, 347(6220):1260419, January 2015. Publisher: American Association for the Advancement of Science.

[171] Mathias Uhlen, Max J. Karlsson, Wen Zhong, Abdellah Tebani, Christian Pou, Jaromir Mikes, Tadepally Lakshmikanth, Björn Forsström, Fredrik Edfors, Jacob Odeberg, Adil Mardinoglu, Cheng Zhang, Kalle von Feilitzen, Jan Mulder, Evelina Sjöstedt, Andreas Hober, Per Oksvold, Martin Zwahlen, Fredrik Ponten, Cecilia Lindskog, Åsa Sivertsson, Linn Fagerberg, and Petter Brodin. A genome-wide transcriptomic analysis of protein-coding genes in human blood cells. *Science*, 366(6472):eaax9198, December 2019. Publisher: American Association for the Advancement of Science.

[172] Yingdong Zhao, Ming-Chung Li, Mariam M. Konaté, Li Chen, Biswajit Das, Chris Karlovich, P. Mickey Williams, Yvonne A. Evrard, James H. Doroshow, and Lisa M. McShane. TPM, FPKM, or Normalized Counts? A Comparative Study of Quantification Measures for the Analysis of RNA-seq Data from the NCI Patient-Derived Models Repository. *Journal of Translational Medicine*, 19(1):269, June 2021.

[173] Joseph Nasser, Drew T. Bergman, Charles P. Fulco, Philine Guckelberger, Benjamin R. Doughty, Tejal A. Patwardhan, Thouis R. Jones, Tung H. Nguyen, Jacob C. Ulirsch, Fritz Lekschas, Kristy Mualim, Heini M. Natri, Elle M. Weeks, Glen Munson, Michael Kane, Helen Y. Kang, Ang Cui, John P. Ray, Thomas M. Eisenhaure, Ryan L. Collins, Kushal Dey, Hanspeter Pfister, Alkes L. Price, Charles B. Epstein, Anshul Kundaje, Ramnik J. Xavier, Mark J. Daly, Hailiang Huang, Hilary K. Finucane, Nir Hacohen, Eric S. Lander, and Jesse M. Engreitz. Genome-wide enhancer maps link risk variants to disease genes. *Nature*, 593(7858):238–243, May 2021. Number: 7858 Publisher: Nature Publishing Group.

[174] Dongkyeong Kim, Hongjoo An, Randall S. Shearer, Mohamed Sharif, Chuandong Fan, Jin-ok Choi, Sun Ryu, and Yungki Park. A principled strategy for mapping enhancers to genes. *Scientific Reports*, 9(1):11043, July 2019. Number: 1 Publisher: Nature Publishing Group.

[175] Andreas R. Gschwind, Kristy S. Mualim, Alireza Karbalayghareh, Maya U. Sheth, Kushal K. Dey, Evelyn Jagoda, Ramil N. Nurtdinov, Wang Xi, Anthony S. Tan, Hank Jones, X. Rosa Ma, David Yao, Joseph Nasser, Žiga Avsec, Benjamin T. James, Muhammad S. Shamim, Neva C. Durand, Suhas S. P. Rao, Ragini Mahajan, Benjamin R. Doughty, Kalina Andreeva, Jacob C. Ulirsch, Kaili Fan, Elizabeth M. Perez, Tri C. Nguyen, David R. Kelley, Hilary K. Finucane, Jill E. Moore, Zhiping Weng, Manolis Kellis, Michael C. Bassik, Alkes L. Price, Michael A. Beer, Roderic Guigó, John A. Stamatoyannopoulos, Erez Lieberman Aiden, William J. Greenleaf,

Christina S. Leslie, Lars M. Steinmetz, Anshul Kundaje, and Jesse M. Engreitz. An encyclopedia of enhancer-gene regulatory interactions in the human genome, November 2023. Pages: 2023.11.09.563812 Section: New Results.

[176] Simon Fishilevich, Ron Nudel, Noa Rappaport, Rotem Hadar, Inbar Plaschkes, Tsippi Iny Stein, Naomi Rosen, Asher Kohn, Michal Twik, Marilyn Safran, Doron Lancet, and Dana Cohen. GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database: The Journal of Biological Databases and Curation*, 2017:bax028, April 2017.

[177] Joannella Morales, Shashikant Pujar, Jane E. Loveland, Alex Astashyn, Ruth Bennett, Andrew Berry, Eric Cox, Claire Davidson, Olga Ermolaeva, Catherine M. Farrell, Reham Fatima, Laurent Gil, Tamara Goldfarb, Jose M. Gonzalez, Diana Haddad, Matthew Hardy, Toby Hunt, John Jackson, Vinita S. Joardar, Michael Kay, Vamsi K. Kodali, Kelly M. McGarvey, Aoife McMahon, Jonathan M. Mudge, Daniel N. Murphy, Michael R. Murphy, Bhanu Rajput, Sanjida H. Rangwala, Lillian D. Riddick, Françoise Thibaud-Nissen, Glen Threadgold, Anjana R. Vatsan, Craig Wallin, David Webb, Paul Flicek, Ewan Birney, Kim D. Pruitt, Adam Frankish, Fiona Cunningham, and Terence D. Murphy. A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. *Nature*, 604(7905):310–315, April 2022. Number: 7905 Publisher: Nature Publishing Group.

[178] Panagiotis Korfiatis, Timothy L. Kline, Daniel H. Lachance, Ian F. Parney, Jan C. Buckner, and Bradley J. Erickson. Residual Deep Convolutional Neural Network Predicts MGMT Methylation Status. *Journal of Digital Imaging*, 30(5):622–628, October 2017.

[179] Scott M Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[180] Mary J. Goldman, Brian Craft, Mim Hastie, Kristupas Repečka, Fran McDade, Akhil Kamath, Ayan Banerjee, Yunhai Luo, Dave Rogers, Angela N.

Brooks, Jingchun Zhu, and David Haussler. Visualizing and interpreting cancer genomics data via the Xena platform. *Nature Biotechnology*, 38(6):675–678, June 2020. Number: 6 Publisher: Nature Publishing Group.

[181] Marian Mellén, Pinar Ayata, and Nathaniel Heintz. 5-hydroxymethylcytosine accumulation in postmitotic neurons results in functional demethylation of expressed genes. *Proceedings of the National Academy of Sciences*, 114(37):E7812–E7821, September 2017. Publisher: Proceedings of the National Academy of Sciences.

[182] Gulfem D. Guler, Yuhong Ning, Chin-Jen Ku, Tierney Phillips, Erin McCarthy, Christopher K. Ellison, Anna Bergamaschi, Francois Collin, Paul Lloyd, Aaron Scott, Michael Antoine, Wendy Wang, Kim Chau, Alan Ashworth, Stephen R. Quake, and Samuel Levy. Detection of early stage pancreatic cancer using 5-hydroxymethylcytosine signatures in circulating cell free DNA. *Nature Communications*, 11(1):5270, October 2020. Number: 1 Publisher: Nature Publishing Group.

[183] Shu Yi Shen, Rajat Singhania, Gordon Fehringer, Ankur Chakravarthy, Michael H. A. Roehrl, Dianne Chadwick, Philip C. Zuzarte, Ayelet Borgida, Ting Ting Wang, Tiantian Li, Olena Kis, Zhen Zhao, Anna Spreafico, Tiago da Silva Medina, Yadon Wang, David Roulois, Ilias Ettayebi, Zhuo Chen, Signy Chow, Tracy Murphy, Andrea Arruda, Grainne M. O'Kane, Jessica Liu, Mark Mansour, John D. McPherson, Catherine O'Brien, Natasha Leighl, Philippe L. Bedard, Neil Fleshner, Geoffrey Liu, Mark D. Minden, Steven Gallinger, Anna Goldenberg, Trevor J. Pugh, Michael M. Hoffman, Scott V. Bratman, Rayjean J. Hung, and Daniel D. De Carvalho. Sensitive tumour detection and classification using plasma cell-free DNA methylomes. *Nature*, 563(7732):579–583, November 2018. Number: 7732 Publisher: Nature Publishing Group.

[184] Jaime Sevilla, Lennart Heim, Anson Ho, Tamay Besiroglu, Marius Hobbhahn, and Pablo Villalobos. Compute Trends Across Three Eras of Machine Learning. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, July 2022. arXiv:2202.05924 [cs].